From a Biased Perspective

- Quasars, Mergers, and Planet-Forming Discs -

Proefschrift

ter verkrijging van
de graad van doctor aan the Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 29 oktober 2025
klokke 13:00 uur

door

Elia Pizzati

geboren te Abano Terme, PD, Italië
in 1997

Promotores:

Prof. dr. J. F. Hennawi

Prof.dr. J. Schaye

Promotiecommissie:

Prof.dr. I.A.G. Snellen

Prof.dr. E. M. Rossi

Prof.dr. M. Volonteri (Institut d'Astrophysique de Paris)

Prof.dr. Z. Haiman (Institute of Science and Technology Austria)

Prof.dr. R. J. Bouwens

Dr. M. Schaller

Cover: Design and Illustration by Monica Rocco

ISBN: 978-94-6510-959-6

Printed by: ProefschriftMaken

An electronic version of this dissertation is available at
http://scholarlypublications.universiteitleiden.nl/.

"lei signora Ph(i)NK_o, quella che in mezzo al chiuso nostro mondo meschino era stata capace d'uno slancio generoso, il primo, "Ragazzi, che tagliatelle vi farei mangiare!", un vero slancio d'amore generale, dando inizio nello stesso momento al concetto di spazio, e allo spazio propriamente detto, e al tempo, e alla gravitazione universale, e all'universo gravitante"
– *Le cosmicomiche*, Italo Calvino

"she, Mrs. Ph(i)Nk_o, she who in the midst of our closed, petty world had been capable of a generous impulse – the very first one – "Boys, the tagliatelle I would make for you!", a true outburst of general love, initiating at the same moment the concept of space and, properly speaking, space itself, and time, and universal gravitation, and the gravitating universe"
– *Cosmicomics*, Italo Calvino

Contents

1 Introduction **1**

1.1 Setting the stage: ΛCDM cosmology and the large-scale structure of our Universe . . . 4

1.1.1 Dark matter halos as the building blocks of cosmic structure formation . . . 7

1.2 Galaxies and their central black holes . . . 9

1.2.1 Black holes: a journey through ten orders of magnitude in the mass spectrum . . . 12

1.2.2 Quasars as tracers of SMBH growth . . . 14

1.2.3 The high-redshift frontier . . . 17

1.3 Observations: the evolution of quasars and SMBHs across cosmic times . . . 19

1.3.1 The quasar luminosity function . . . 20

1.3.2 Quasar clustering and the duty cycle of quasars . . . 21

1.3.3 SMBH mergers and gravitational waves . . . 24

1.3.4 New challenges in the JWST era: the nature of "little red dots" and other broad-line AGN . . . 26

1.4 Theoretical models: key uncertainties and future directions . 29

1.5 This thesis . . . 32

2 Revisiting the extreme clustering of $z \approx 4$ quasars with large volume cosmological simulations **37**

2.1 Introduction . . . 38

2.2 Methods . . . 42

2.2.1 The conditional luminosity function . . . 42

2.2.1.1 The quasar auto-correlation function . . . 45

2.2.1.2 Halo occupation distribution and duty cycle 46

2.2.2 Dark matter only simulation setup . . . 47

2.2.2.1 Fitting the halo mass function . . . 51

2.2.2.2 Obtaining the cross-correlation functions . . 52

2.3 Data-model comparison . . . 54

2.3.1 Overview of observational data . . . 54

2.3.2 Likelihood functions . . . 56

2.4 Results . . . 58

2.4.1 Analysis at $z \approx 4$ . . . 58
2.4.1.1 MCMC analysis . . . 63
2.4.2 Comparison with $z \approx 2.5$ . . . 64
2.5 Discussion . . . 68
2.5.1 Implications for quasars' physical properties . . . 69
2.5.1.1 Black hole mass and accretion efficiency . . . 69
2.5.1.2 Quasar lifetime and the growth of high-z black holes . . . 71
2.5.2 Comparison with previous work . . . 74
2.5.3 Caveats and final remarks . . . 77
2.6 Summary . . . 79
2.A Appendix: Obtaining the quasar auto-correlation from the halo cross-correlation functions . . . 82
2.B Appendix: Fitting the cross-correlation terms from simulations 84
2.C Appendix: Halo mass function and correlation functions for redshift $z = 2.5$ . . . 85

3 A unified model for the clustering of quasars and galaxies at $z \approx 6$ 91
3.1 Introduction . . . 92
3.2 Methods . . . 96
3.2.1 Quasar and galaxy population models . . . 96
3.2.2 Simulation setup . . . 99
3.2.2.1 Extending the suite of FLAMINGO runs: FLAMINGO-10k . . . 100
3.2.2.2 Obtaining the sub-halo catalogue with HBT+ 101
3.2.2.3 A simulation-based analytical description of halo properties . . . 102
3.3 Data-Model comparison . . . 104
3.3.1 Overview of observational data . . . 106
3.3.2 Parameter inference . . . 107
3.4 Results . . . 111
3.4.1 The quasar luminosity-halo mass relation and the host halos of quasars at $z \approx 6$ . . . 113
3.4.2 Characterizing the properties of [OIII] emitters . . . 115
3.5 Discussion . . . 119
3.5.1 Quasar properties across cosmic time . . . 119
3.5.2 The quasar duty cycle and SMBH growth . . . 124
3.6 Summary . . . 125
3.A Appendix: Details on the conditional luminosity function framework . . . 128

3.B Appendix: Results for the fitting of the halo cross-correlation functions . . . 130
3.C Appendix: Interpreting the auto-correlation measurements of $z \approx 6$ quasars . . . 133
3.D Appendix: Quasar-host halo masses with a uniform luminosity threshold . . . 135

4 "Little red dots" cannot reside in the same dark matter halos as comparably luminous unobscured quasars 139
4.1 Introduction . . . 140
4.2 The staggeringly high abundance of UV-obscured AGN implied by little red dots . . . 145
4.3 Little red dots and UV-selected quasars: do they belong to the same population? . . . 151
4.3.1 The host dark matter halos and duty cycles of high-z unobscured quasars and their luminosity dependence . 153
4.3.2 Connecting the UV-luminous duty cycle to the AGN obscured population . . . 159
4.4 The host mass and duty cycle of little red dots: a mock analysis 163
4.5 Discussion and summary . . . 167

5 Tracing individual black hole growth histories and quasar lightcurves in an N-body Universe 173
5.1 Introduction . . . 174
5.2 Methods . . . 178
5.2.1 Extracting halo mass histories and merger trees from the FLAMINGO simulation . . . 179
5.2.1.1 Subhalo masses and specific halo accretion rates . . . 180
5.2.1.2 Construction of the merger tree catalogs . . 182
5.2.2 Modeling SMBH and quasar evolution . . . 183
5.2.2.1 Black hole initialization . . . 185
5.2.2.2 Black hole mergers . . . 185
5.2.2.3 Black hole accretion and quasar radiation . . 186
5.2.3 Overview of the observational constraints . . . 190
5.2.4 Fiducial model and parameter inference . . . 194
5.3 Results . . . 197
5.3.1 The buildup of supermassive black holes across cosmic history . . . 197
5.3.2 Comparison with quasar observables . . . 201
5.3.3 Implications for SMBH growth and scaling relations . 208

5.3.3.1 The black hole mass-halo mass relation across cosmic history . . . 208
5.3.3.2 The coherence timescale of the accretion process . . . 211
5.3.3.3 The relative role of mergers and accretion . . . 213
5.4 Discussion and summary . . . 215

6 Towards inference of overlapping grav. wave signals 223
6.1 Introduction . . . 224
6.2 Number of overlapping signals . . . 227
6.2.1 Overlapping signals of the same family . . . 228
6.2.2 Overlapping signals from two different families . . . 229
6.3 Covariance among overlapping signals . . . 229
6.4 Bayesian inference of overlapping signals . . . 233
6.4.1 Choice of signal families . . . 234
6.4.2 Setting up Bayesian inference runs . . . 236
6.4.3 Bayesian priors . . . 237
6.4.4 Results . . . 238
6.4.4.1 Dependence on the luminosity distance . . . 240
6.4.4.2 Dependence on the phase . . . 242
6.5 Discussion and Outlook . . . 243

7 Constraining turbulence in protoplanetary discs using the gap contrast: an application to the DSHARP sample 247
7.1 Introduction . . . 248
7.2 Methods . . . 250
7.2.1 Disc structure . . . 251
7.2.2 Inferring the disc surface density . . . 252
7.2.3 Analysis of the gaps filling effect on the minor axis . . . 255
7.3 DSHARP data sample . . . 256
7.4 Results . . . 256
7.4.1 GW Lup as a case study . . . 258
7.4.2 Overview of the other systems . . . 260
7.4.2.1 DoAr 25 . . . 263
7.4.2.2 Elias 24 . . . 264
7.4.2.3 AS 209 . . . 264
7.4.2.4 HD 163296 . . . 265
7.4.2.5 MY Lup . . . 266
7.5 Discussion . . . 266
7.5.1 Relative dust and gas scale heights . . . 267
7.5.2 Implications for turbulence . . . 268

7.5.3 When does the method yield constraints on the scale height? . . . 269
7.5.4 Caveats . . . 271
7.6 Summary . . . 272
7.A Appendix: Discs with no constraints . . . 274
7.B Appendix: Convergence along the major axis and emission maps . . . 274

Bibliography **279**

English summary **297**

Nederlandse samenvatting **301**

Riassunto in italiano **305**

List of Publications **309**

Curriculum Vitae **313**

Acknowledgements **315**

1 Introduction

Being born at the end of the last century, it's difficult for me to fully appreciate just how dramatic – and how recent – the progress in our understanding of the Universe has been. For millennia, civilizations devised myths to explain the origins of the cosmos and its fate. Today, however, we possess a well-tested theoretical framework that describes the evolution of the Universe from about one second after the Big Bang to the present day and beyond. Known as the standard model of cosmology, or the ΛCDM model, this framework is rooted in Einstein's theory of general relativity and has been refined through a series of paradigm shifts, some of the most consequential occurring only in the past few decades.

Much like the Standard Model of particle physics, ΛCDM is powerful precisely because it explains a vast range of observations using only a small number of parameters. It has withstood numerous experimental tests and challenges. While it is not the final word – offering limited insight into the true nature of dark matter, dark energy, or the physics of inflation – it remains the most robust and predictive model for describing the large-scale evolution of the Universe.

ΛCDM provides the backbone: a model of the Universe on the largest scales. On top of this framework, however, lies the rich tapestry of structure formation: the growth of stars, galaxies, and black holes. Research in physical cosmology today is often split along two broad directions. On one side is the pursuit of ever more precise tests of the ΛCDM model, a field now driven by percent-level measurements that seek signs of new physics in the smallest deviations from theoretical predictions. This often relies on observations of the largest cosmic scales, where the imprint of complex astrophysical processes is negligible or can be marginalized effectively. On the other side lies the study of those very astrophysical processes – nonlinear, chaotic, and often poorly understood. This is not a percent-level science, but an order-of-magnitude one. Yet, it is precisely this limited understanding of galaxy evolution that has made the field ripe for discovery, and indeed, some of the most transformative progress in recent years has come from this domain.

A crucial driver of this recent progress has been the advent of new observatories capable of probing ever earlier cosmic times. By pushing the limits of our observations to higher redshifts[1], these instruments are extending the frontiers of our knowledge and opening up new regimes for discovery. The most prominent recent example is the James Webb Space

[1] The cosmological redshift, z, is a measure of cosmic time: higher redshifts correspond to earlier times in the history of the Universe.

Telescope (JWST), launched in 2021. With its unprecedented sensitivity and resolution, JWST is already reshaping our understanding of the early Universe by revealing surprising properties of the first galaxies and black holes and challenging established theoretical models. It has been an instructive experience to witness this chaotic yet inspiring transformative process taking place during my Ph.D. – progress that also shapes the core of this thesis.

Among the most important findings in the context of galaxy formation is the central role played by supermassive black holes (SMBHs). Initially proposed in the 1960s to explain the immense energy output of quasars, SMBHs quickly became a cornerstone of active galactic nuclei (AGN) theory. By the 1990s, high-resolution observations showed that SMBHs are ubiquitous in the Universe, inhabiting all massive galaxies even in the absence of AGN activity. Tight empirical correlations were discovered between the mass of SMBHs and several properties of the host galaxies, suggesting a scenario of coevolution, or, at the very least, a deep physical connection between black hole growth and galaxy evolution. More recently, the detection of luminous quasars at redshifts beyond $z > 6$ has shown that billion-solar-mass black holes were already in place less than a billion years after the Big Bang, posing significant challenges to our understanding of early black hole growth.

This thesis explores these themes – among others – through a focused investigation of SMBHs as traced by the properties of quasars, particularly in the high-redshift Universe where the earliest stages of SMBH and quasar evolution unfold. Chapters 2–5 are dedicated to this topic. The central idea driving these chapters is that the ΛCDM model provides a robust cosmological framework for describing the large-scale structure and its evolution across cosmic time. By building upon this foundation, we can construct models to understand how SMBHs and quasars form, grow, and evolve, and connect them to observations across a range of redshifts. The overarching goal is to determine how complex astrophysical processes – such as black hole accretion and quasar activity – can be consistently embedded within the standard cosmological paradigm, using a phenomenologically driven approach that bridges theory and data.

Chapters 6 and 7 focus on distinct research directions that fall outside the main scope of this thesis but reflect other projects I pursued before and during my Ph.D., in collaboration with researchers from other subfields. Chapter 6 shifts attention to the opposite end of the black hole mass spectrum: stellar-mass black holes, whose mergers have been detected by gravitational wave observatories such as LIGO, Virgo, and KAGRA. This chapter presents a first step toward addressing a key challenge for the next generation of gravitational wave detectors – expected to operate in the 2030s – which will be so sensitive that overlapping signals in the time domain may complicate the inference of source parameters. Our work quantifies and assesses the impact of this overlap on parameter estimation.

Chapter 7 explores a different class of astrophysical discs – not those around SMBHs, but the protoplanetary discs surrounding young stars, which are the birthplaces of planetary systems. These structures have been extensively studied with the Atacama Large Millimeter/submillimeter Array (ALMA), which provides high-resolution observations in the far infrared. In this chapter, we use ALMA data to infer the vertical structure of protoplanetary discs, shedding light on the early stages of planet formation.

While this introduction focuses primarily on supermassive black holes and extragalactic astrophysics, a complete overview of the contents of all chapters is provided in Section 1.5.

Why this title

What links bright quasars in galaxies far, far away to mergers of stellar-mass black holes in the nearby Universe – detected through their gravitational-wave emission – to the protoplanetary discs we observe in our solar neighbourhood? I asked myself this question when searching for a title for this thesis. While many connections can be drawn, one in particular stood out to me: the idea that, in different ways, this thesis revolves around the concept of *bias*.

Chapter 1 opens by discussing how quasars at $z \approx 4$ appear to be among the most biased tracers of structure in the early Universe. There, we are talking about *cosmological bias* – a concept introduced in Sec. 1.1. In this context, bias quantifies how the spatial distribution of certain astrophysical objects, like quasars, relates to the underlying distribution of dark matter. Because they tend to reside in massive structures, quasars are more clustered than the matter field as a whole, and thus are said to be "biased" tracers of the large-scale distribution of matter. Understanding how different populations of quasars and galaxies trace this distribution is central to embedding black hole and galaxy evolution within a cosmological framework.

Cosmological bias is a somewhat niche but well-defined concept. Later in the thesis, a different kind of bias takes the stage – one that is more familiar to anyone working in data analysis or modeling. Chapter 6 deals with *bias in parameter inference*: how assumptions, modelling choices, and incomplete information can systematically skew the results we extract from data. In this specific case, we assess the bias that can arise from the overlap of multiple gravitational wave signals in the time domain.

Finally, there's a broader, more implicit sense in which bias plays a role: the idea that our window on the Universe is inevitably a biased point of observation. This *observational bias* lies at the core of astronomy as a science. As we show in Chapter 7, however, this bias can also be turned into an advantage. Protoplanetary discs, which are approximately azimuthally symmetric in three dimensions, appear as ellipses on the sky due to projection effects. These same effects distort the apparent shape of substructures – such as the rings and gaps that often characterize discs – breaking their

symmetry in a predictable way that depends on the disc's inclination and intrinsic morphology. As a result, our particular vantage point allows us to constrain the three-dimensional structure of discs, using the bias introduced by projection as a diagnostic tool rather than a limitation.

"From a biased perspective" also reflects my own path through astrophysics. It has been a biased one – shaped by curiosity, but also by chance encounters, guidance from mentors, and the particular set of tools and questions I ended up gravitating towards. There is no single pattern or overarching plan, no carefully laid-out roadmap guiding the journey.

Bias, in its many forms, is something we must acknowledge – whether we aim to model it, correct for it, or simply be aware of it. It shapes what we see, how we interpret it, and what we conclude. But it also reflects who we are: our interests, our choices, our perspective on the Universe. This thesis is one such perspective – a biased one, certainly, but hopefully one worth telling.

1.1 Setting the stage: ΛCDM cosmology and the large-scale structure of our Universe

The ΛCDM model is founded on the "cosmological principle", which posits that, on sufficiently large scales, the Universe is both homogeneous (the same everywhere) and isotropic (the same in all directions). This principle allows a great simplification of Einstein's field equations in general relativity, yielding dynamic solutions that describe a Universe whose overall scale evolves with time, either expanding or contracting according to a scale factor $a(t)$. The corresponding spacetime geometry is captured by the Friedmann-Lemaître-Robertson-Walker (FLRW) metric, which underpins the standard cosmological model. The Friedmann equations relate the dynamics of the scale factor, $a(t)$, to the energy content of the Universe (Friedmann 1922, 1924).

Observational support for this framework came with Edwin Hubble's discovery that distant galaxies exhibit a systematic redshift, indicating that the Universe is indeed expanding (Hubble 1929). This interpretation was further reinforced by Georges Lemaître, who proposed that such expansion implies a finite age and a primordial, hot, and dense origin – a concept that would later be termed the "Big Bang" (Lemaître 1931). In the context of modern cosmology, this singular beginning marks the onset of cosmic time and the starting point for the formation and evolution of all known structures in the Universe.

A cornerstone of the ΛCDM model is the realization that the energy content of the Universe is not composed solely of ordinary (baryonic) matter. Rather, baryons account for only $\approx 5\%$ of the total energy density of

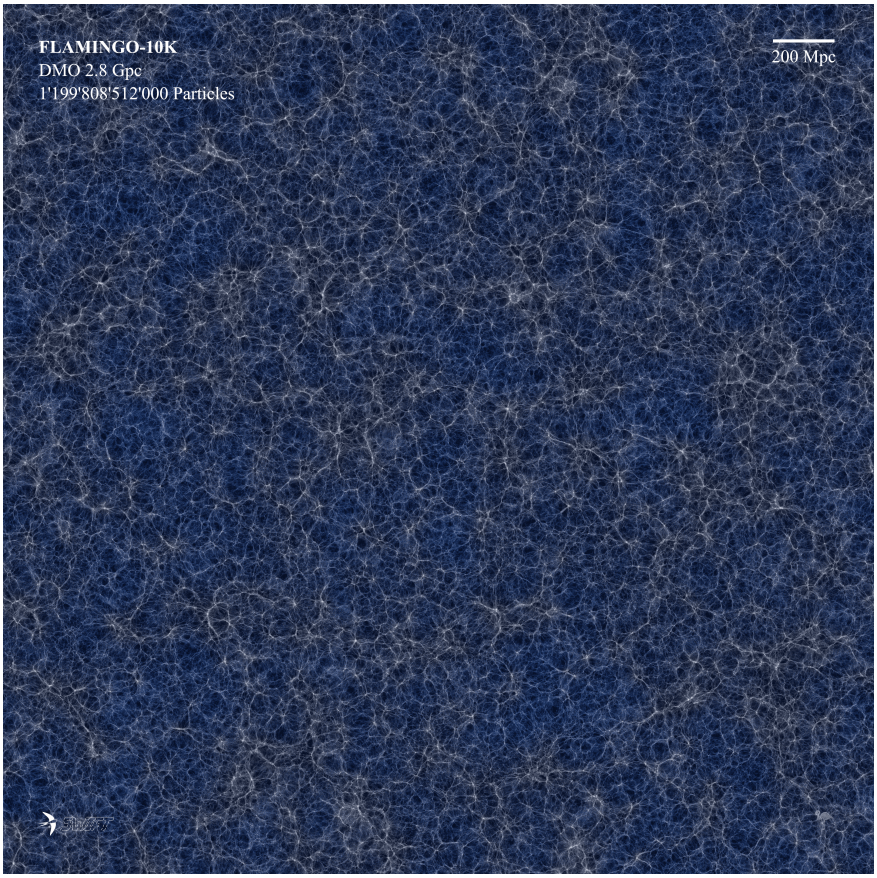

Figure 1.1: The large-scale structure of the Universe as predicted by the FLAMINGO-10k dark-matter-only (DMO) cosmological simulation (Schaller et al., in prep.; see also Chapter 3). This simulation was run by evolving over one trillion particles within a box measuring 2.8 cGpc per side, making it one of the largest simulations ever performed. A version of this image was turned into a puzzle – courtesy of Matthieu Schaller – and shared with the Leiden Observatory community. In assembling it, Ph.D. students at Leiden were reminded of Einstein's cosmological principle the hard way: the Universe gets surprisingly boring when you zoom out too far.

the Universe at the present epoch. The remaining components include dark energy (Λ, which can be associated with the cosmological constant in Einstein's field equations), cold dark matter (CDM), radiation, and neutrinos. While dark energy drives the accelerated expansion of the Universe at late times, cold dark matter – a non-luminous, collisionless form of matter – is essential for the formation and growth of cosmic structures. Although its precise nature remains unknown and direct detection has not yet been achieved, the gravitational influence of dark matter is indispensable for reconciling theoretical predictions with a wide range of astrophysical observations, including the dynamics of galaxies, gravitational lensing, and the large-scale clustering of matter.

Even though the Friedmann equations assume a perfectly smooth Universe, we know that in reality, it contains small Gaussian density perturbations. The origin of these fluctuations is still not fully understood – likely, they arise from quantum fluctuations in the very early Universe, which were stretched to macroscopic scales during a brief period of cosmic inflation. But we know for a fact that these fluctuations exist: we see them imprinted in the Cosmic Microwave Background (CMB), which offers a direct snapshot of the Universe at a redshift $z \approx 1100$. The exquisite measurements of the CMB – particularly by missions such as WMAP and Planck (Spergel et al. 2007; Planck Collaboration et al. 2014) – provide strong evidence for the statistical properties of these primordial fluctuations.

These small overdensities in the primordial density field act as the seeds from which all cosmic structures emerge. In the early Universe, these perturbations evolve linearly: fluctuations at different scales grow independently and proportionally to a common, time-dependent growth factor. Within this regime, linear perturbation theory provides an accurate analytic description of the evolution of the matter density field (e.g., Peebles 1980). As cosmic time progresses, however, overdensities grow under the influence of gravity and eventually reach the threshold at which linear theory ceases to be valid. This transition marks the onset of the non-linear regime of structure formation. Gravitational collapse proceeds anisotropically – typically beginning along the shortest axis of a perturbation – giving rise to the complex filamentary pattern known as the cosmic web (see Fig. 1.1; Bond et al. 1996). When matter collapses along all three spatial dimensions, it forms bound, virialized structures known as *halos*. These halos are made of dark matter, which collapses earlier than their baryonic counterparts due to its collisionless nature. Once halos form, they act as gravitational wells that attract baryonic gas, allowing it to cool, condense, and eventually form stars and galaxies (Barkana & Loeb 2001; Dayal & Ferrara 2018).

To study the non-linear regime of structure formation, cosmologists employ N-body simulations, which numerically evolve a system of particles under their mutual gravitational interactions. Since dark matter dominates the mass budget of the Universe and drives the formation of structures, these

simulations typically include only dark matter (and, in some cases, massive neutrinos). They are therefore commonly referred to as dark-matter-only simulations. In contrast, baryonic processes – such as gas dynamics, radiative cooling, star formation, and feedback from supernovae and AGN – introduce considerable physical complexity and are associated with significant modeling uncertainties (Vogelsberger et al. 2020). These effects become prominent mainly at the scale of dark matter halos, where non-gravitational forces play a critical role. On larger scales, however, the influence of baryons is subdominant, allowing the large-scale structure of the Universe to be accurately described using gravity alone.

Since the pioneering efforts in the early 1980s (e.g., Efstathiou et al. 1985; Davis et al. 1985), the field of N-body simulations has evolved significantly, driven by advances in algorithms and computational capabilities (Angulo & Hahn 2022). Modern simulations are able to capture the formation and evolution of structures with remarkable accuracy, extending from ∼Gpc scales down to the ∼kpc scale of subhalos. Figure 1.1 presents an example of the projected matter density field from the FLAMINGO-10k cosmological simulation (Schaller et al., in prep.). With more than 10^{12} dark matter (CDM) particles evolved in a 2.8 cGpc box, it represents one of the largest simulations ever run.

1.1.1 Dark matter halos as the building blocks of cosmic structure formation

By acting as gravitational wells that attract baryons, dark matter halos are the fundamental environments within which visible structures in the Universe originate. Infalling gas cools radiatively and condenses at the centre of halos, eventually giving rise to stars, galaxies, and supermassive black holes. Consequently, there exists a close connection between the hierarchical assembly of dark matter halos and the formation and evolution of galaxies. This correspondence forms the basis of theoretical frameworks such as semi-analytic models (SAMs) and semi-empirical approaches, which model galaxy formation and evolution by tracking the distribution and merging histories of halos and subhalos (i.e., satellite halos contained within a larger halo) across cosmic time (Somerville & Davé 2015; Lapi et al. 2025).

The distribution and evolution of (sub)halos can be extracted from N-body cosmological simulations using (sub)halo finder algorithms. These algorithms analyze the particle data output from simulations to identify gravitationally bound structures. Typically, halo finders first locate candidate halos based on density peaks or groups of particles connected in configuration, phase, or history space, and then apply an unbinding procedure to remove particles that are not gravitationally bound (Onions et al. 2012; Forouhar Moreno et al. 2025).

An alternative to explicitly identifying halos in simulations is provided by the halo model (e.g., Cooray & Sheth 2002). This empirical framework assumes that all matter resides in dark matter halos and uses analytic prescriptions for halo properties, abundances, and spatial distributions to statistically describe the large-scale matter field. Rather than tracking individual halos, the halo model predicts ensemble-averaged quantities – such as correlation functions or power spectra – by extending linear theory using analytic prescriptions (Asgari et al. 2023). While effective at quasi-linear scales ($r \gtrsim 10\,\mathrm{cMpc}$), the halo model becomes increasingly inaccurate at small scales, high redshifts, and large halo masses, due to its simplified assumptions about halo profiles, substructure, and non-linear effects (e.g., Mead & Verde 2021). These limitations make it unsuitable for the regimes explored in this thesis. For this reason, we rely on large-volume N-body simulations to extract accurate halo statistics in Chapters 2–5.

The most fundamental statistic describing the halo population is the halo mass function (HMF), which quantifies the comoving number density of halos as a function of mass. In its simplest analytical form, the HMF can be derived using the Press–Schechter formalism (Press & Schechter 1974). This approach assumes that halos form from regions in the initial density field where the linearly extrapolated density contrast, smoothed on some scale, exceeds a critical threshold for collapse. This allows one to relate the abundance of halos to the statistical properties of the initial Gaussian density field. While the Press–Schechter model captures the essential physics of hierarchical collapse, it relies on simplifying assumptions and underestimates halo abundances at both the low- and high-mass ends. Modern N-body simulations offer precise empirical descriptions of the HMF over a wide range of halo masses and redshifts, accounting for the full non-linear dynamics of structure formation (e.g., Tinker et al. 2008; Bocquet et al. 2016).

Another key property of halos is their spatial clustering. Because halos form from peaks in the initial density field, their distribution is biased relative to the underlying matter distribution. This phenomenon, known as cosmological halo bias, is scale-dependent in general, but on large (linear) scales, it is well described by a mass- and redshift-dependent bias factor.

The origin of halo bias can be understood through the statistics of Gaussian random fields. Rarer peaks – associated with more massive halos – require the constructive interference of more Fourier modes and are thus more clustered than typical regions. More quantitatively, a halo of mass M at redshift z corresponds to a peak in the smoothed linear density field whose height is given by $\nu(M, z) = \delta_c(z)/\sigma(M)$, where $\delta_c(z)$ is the critical linear overdensity for collapse, and $\sigma(M)$ is the standard deviation of the density field smoothed on the scale corresponding to mass M. Higher peak heights ($\nu \gg 1$) correspond to rarer, more strongly clustered halos. This implies that halo clustering strength increases with both halo mass and redshift. For instance, at fixed mass, halos forming at earlier times must come from

higher peaks, since the growth factor is smaller and fluctuations must be intrinsically larger to collapse by that time. Similarly, at fixed redshift, more massive halos form from rarer, higher peaks, and are therefore more strongly biased.

This connection between halo clustering and mass enables a powerful technique in galaxy formation studies: by comparing the observed clustering of a galaxy population to theoretical predictions for the clustering of halos, one can infer the typical mass of the halos that host those galaxies (e.g., Mo & White 1996). This provides crucial insight into the connection between galaxies and their large-scale environments. In Chapters 2-5, we will apply the same idea to connect quasars to halos and galaxies by using their measured clustering properties (see Sec. 1.3.2).

1.2 Galaxies and their central black holes

In 1925, Edwin Hubble used the period-luminosity relation of Cepheid variables – originally discovered by Henrietta Leavitt – to demonstrate that the Andromeda Nebula lies well beyond the boundaries of the Milky Way (Hubble 1925). This discovery resolved the "Great Debate" and marked the beginning of extragalactic astronomy: the Universe consists not just of stars within our own galaxy, but of countless other galaxies scattered across cosmic space (Trimble 1995).

One century later, our view of the cosmos has dramatically expanded. We now routinely observe hundreds of billions of galaxies, tracking their formation and evolution from the earliest epochs (up to $z \approx 14$, Carniani et al. 2024) to the present day. The classical picture of galaxies as isolated "island universes", a concept popularized by Immanuel Kant, has been replaced by a far more dynamic one. Galaxies are not self-contained or static – they are complex, interconnected ecosystems, shaped by both their internal processes and their interactions with the surrounding cosmic environment.

In recent years, theoretical models of galaxy formation have been complemented by the development of large-scale hydrodynamical simulations, which now serve as key tools for studying galaxy evolution in a cosmological setting. Landmark projects such as Illustris (Vogelsberger et al. 2014), EAGLE (Schaye et al. 2015), and IllustrisTNG (Nelson et al. 2019) have demonstrated that it is possible to reproduce a broad range of observed galaxy properties across cosmic time within large cosmological volumes ($100 - 300\,\mathrm{cMpc}$). These simulations go beyond N-body models by solving the coupled equations of gravity and hydrodynamics for baryonic matter, while also incorporating subgrid models for key processes such as star formation and feedback from supernovae and AGN. With the progressive refinement of such models, simulations have reached a stage where they can successfully reproduce numerous observables, including the stellar mass

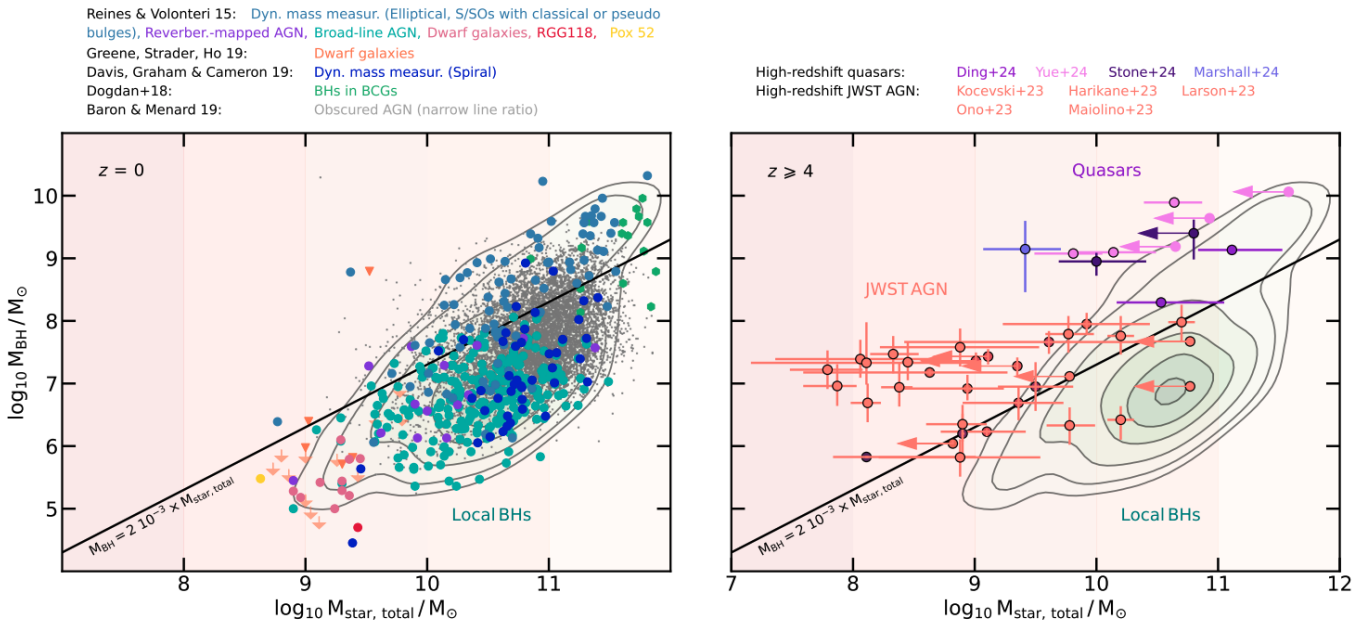

Figure 1.2: *Left:* Black hole mass–stellar mass relation in the local Universe, illustrating the overall connection between galaxy growth and the assembly of their central supermassive black holes (SMBHs). SMBH masses are derived using a range of observational techniques, including dynamical modeling, reverberation mapping, and virial estimates from broad emission lines. Different galaxy types and measurement methods are indicated with distinct colors and markers (measurements are taken from Reines & Volonteri 2015; Greene et al. 2020; Davis et al. 2019; Bogdán et al. 2018; Baron & Ménard 2019). *Right:* Same as the left panel, but for the high-redshift Universe ($z > 4$). Sources are divided between quasars (measurements from Ding et al. 2023; Stone et al. 2023; Yue et al. 2024a; Marshall et al. 2023) and JWST-discovered AGN candidates (see Sec. 1.3.4; Kocevski et al. 2024; Maiolino et al. 2024; Harikane et al. 2023; Ono et al. 2023; Larson et al. 2023). The contours represent the distribution of systems in the local Universe and are shown as a reference. Figure taken from Alexander et al. (2025).

function, star formation rates, morphologies, sizes, colors, and the spatial clustering of galaxies (Vogelsberger et al. 2020; Crain & van de Voort 2023).

One of the most significant insights to emerge from both theoretical models and cosmological hydrodynamical simulations is the central role played by supermassive black holes (SMBHs) in shaping the evolution of their host galaxies. Residing at galactic centers, SMBHs are far from being passive end-products of galaxy formation. Instead, they exert a profound influence on their large-scale environment through energetic feedback processes. In particular, AGN and quasar feedback – driven by gas accretion onto SMBHs – plays a pivotal role in regulating star formation, heating and ejecting gas, and ultimately quenching the growth of massive galaxies (Croton et al. 2006; Hopkins et al. 2006; Sijacki et al. 2007; Booth & Schaye 2009).

These feedback mechanisms are now recognized as essential components in explaining a wide range of observed galaxy properties. They help resolve long-standing discrepancies between theoretical predictions and observations, such as the cutoff in the high-mass end of the stellar mass function, the color bimodality of galaxies, and the existence of massive quiescent systems at relatively early cosmic epochs (Sazonov et al. 2005; Fabian 2012; Bower et al. 2006; Somerville & Davé 2015). By linking small-scale black hole accretion physics to large-scale galaxy evolution, AGN feedback has become a cornerstone of modern galaxy formation theory.

Perhaps the clearest observational evidence for the co-evolution of galaxies and their central black holes is the existence of empirical correlations between SMBH mass and several key galaxy properties – such as bulge mass, stellar mass, and circular velocity (Fig. 1.2, left panel). First identified about two decades ago (Magorrian et al. 1998; Gebhardt et al. 2000; Ferrarese & Merritt 2000), these relations imply that the assembly of SMBHs and galaxies is not independent, but regulated by coupled physical processes. As such, they provide key constraints for both semi-analytic models and cosmological simulations, informing prescriptions for black hole seeding, growth, and feedback (Sec. 1.2.2).

At the same time, the origin of these relations, particularly their emergence and evolution at high redshift, remains an open question and a key focus of ongoing research (Volonteri et al. 2021; Greene et al. 2020). Probing how these relations evolved across cosmic time can reveal when and how the coupling between SMBHs and their host galaxies was established, and what are the key drivers of this process. Recent high-redshift observations from JWST have opened a new observational window into this regime. Fig. 1.2 (right panel) presents a compilation of SMBH–host galaxy mass measurements from AGN and quasars at $z > 4$, offering early constraints on the redshift evolution of the scaling relation. Whether the normalization of this relation increases with redshift – suggesting that SMBHs outpace their hosts in early growth – or remains constant, is still a subject of considerable debate (Pacucci et al. 2023; Li et al. 2025b).

In summary, the last few decades have marked a paradigm shift in our understanding of black holes: from abstract mathematical solutions of Einstein's equations to fundamental agents in shaping galaxies and the large-scale structure of the Universe.

1.2.1 Black holes: a journey through ten orders of magnitude in the mass spectrum

The history of black holes as astrophysical objects begins with one of the most remarkable discoveries in observational astronomy: quasars. In 1963, Maarten Schmidt investigated the radio source 3C 273 and realized that its optical spectrum contained redshifted hydrogen emission lines (Schmidt 1963). According to Hubble's law, this redshift implied a cosmological distance of several hundred cMpc, placing it far beyond the local Universe. Given this distance, its observed flux corresponded to a luminosity exceeding that of entire galaxies. Moreover, the rapid variability of its emission constrained the size of the emitting region to less than a parsec, implying an extraordinarily compact and dense energy source.

To account for such features, Edwin Salpeter and Yakov Zeldovich independently proposed in 1964 that the energy source powering quasars must be gravitational accretion of matter onto a massive, compact object – what we now call a supermassive black hole (Salpeter 1964; Zel'dovich & Novikov 1967). Donald Lynden-Bell further developed this idea in 1969, arguing that the infalling material would form a rotating disk, funnelling into what he described as a "Schwarzschild throat" (Lynden-Bell 1969). He went on to suggest that inactive galactic nuclei are simply the fossil remnants of once-luminous quasars, now harboring SMBHs at their cores.

Additional dynamical evidence began to support the existence of SMBHs. In 1970, Wolfe and Burbidge showed that the large stellar velocity dispersions observed in elliptical galaxy nuclei required a mass concentration far exceeding what could be attributed to normal stars (Wolfe & Burbidge 1970). They concluded that a central black hole as massive as $\sim 10^{10}\,\mathrm{M}_\odot$, or a swarm of smaller black holes, could account for the data. The first concrete dynamical detection of such a massive dark object came in 1978 in the galaxy M87, where the core was inferred to host a $\sim 5 \times 10^9\,\mathrm{M}_\odot$ black hole (Sargent et al. 1978). Similar measurements in other galaxies soon followed (e.g., Kormendy 1988).

Closer to home, Lynden-Bell and Rees hypothesized in 1971 that the Milky Way's center should host a massive black hole (Lynden-Bell & Rees 1971). This idea gained traction after the discovery of the compact, bright radio source Sagittarius A* (Sgr A*) in 1974 by Balick & Brown (1974). Follow-up infrared observations over the following decades measured the orbits of individual stars near the Galactic Center with exquisite precision (Eckart & Genzel 1996; Ghez et al. 2008). These showed that Sgr A* must

contain a mass of $\sim 4 \times 10^6\,M_\odot$ confined within a region smaller than the Solar System – evidence so compelling that it earned the 2020 Nobel Prize in Physics for Genzel, Ghez, and Penrose.

Today, it is well established that SMBHs are ubiquitous in the local Universe. Observational surveys show that essentially all galaxies with a bulge component host a central SMBH (Kormendy & Ho 2013). In recent years, the Event Horizon Telescope has provided even more direct evidence: the first resolved images of the event horizons of two SMBHs, in M87 (Event Horizon Telescope Collaboration et al. 2019) and our own Milky Way (Event Horizon Telescope Collaboration et al. 2022).

In parallel to the discovery of quasars and the growing realization that SMBHs reside in galactic nuclei, a different class of black holes was being uncovered with the rise of X-ray astronomy. In the early 1970s, observations with balloon-borne and satellite-based detectors revealed bright X-ray sources in the Milky Way. Among the most notable was Cygnus X-1, whose X-ray variability and association with a massive O-type star pointed to the presence of an unseen, compact companion. Detailed dynamical studies confirmed that the mass of this dark object exceeded the theoretical limit for a neutron star, providing the first compelling evidence for a stellar-mass black hole (Bolton 1972). These black holes are now understood to form as the end products of massive stellar evolution, when the core of a massive star collapses under its own gravity after exhausting its nuclear fuel.

While a handful of stellar-mass black holes were known from X-ray binaries in the late 20th century, the true diversity and abundance of this population remained elusive until the advent of gravitational wave astronomy. Beginning in 2015 with the landmark detection of GW150914, the LIGO and Virgo observatories have opened a new window onto the Universe, directly detecting the mergers of binary black hole systems through their gravitational wave emission (Abbott et al. 2016a). These discoveries unveiled a surprising population of stellar-mass black holes, with masses ranging from a few to over 100 solar masses – challenging preexisting models of stellar evolution and compact object formation (Abbott et al. 2019a). The growing catalog of gravitational wave events now offers an independent, dynamical probe of black hole demographics, complementing electromagnetic observations and revealing regions of parameter space previously inaccessible.

Taken together, black holes span over ten orders of magnitude in mass, from a few solar masses to tens of billions. While solid observational evidence remains limited to the regimes of stellar-mass ($\sim 1 - 100\,M_\odot$) and supermassive ($\sim 10^6 - 10^{10}\,M_\odot$) black holes, the mass distribution of these systems is thought to form a continuum, shaped by diverse evolutionary and growth pathways and possibly by multiple formation channels. The elusive population of intermediate-mass black holes (IMBHs; $\sim 10^2 - 10^5\,M_\odot$) remains poorly constrained observationally, but evidence for their existence is gradually accumulating. This includes the detection of very massive mergers

of stellar-mass black holes (The LIGO Scientific Collaboration et al. 2025), the identification of candidate low-luminosity AGN potentially powered by IMBHs (Greene et al. 2020), and alternative signatures such as tidal disruption events (Zhang et al. 2025) or dynamical studies of dense stellar systems, with recent claims involving systems like ω Centauri (Häberle et al. 2024).

1.2.2 Quasars as tracers of SMBH growth

The journey of black holes across the mass spectrum remains largely mysterious – particularly at the low-mass end, where formation pathways and early evolution are still poorly constrained by observations. At the opposite end of the spectrum, however, the formation and evolution of SMBHs is illuminated by a simple and elegant argument first articulated by Soltan (1982). The "Soltan argument" asserts that the same accretion processes powering luminous quasars naturally account for the buildup of SMBH mass over cosmic time.

The key idea is as follows: when gas is accreted onto a black hole at a rate $\dot{M}_{\rm acc}$, a fraction ϵ (known as *radiative efficiency*) of its rest-mass energy is converted into radiation. This results in a bolometric luminosity given by:

$$L_{\rm bol} = \epsilon \dot{M}_{\rm acc} c^2. \tag{1.1}$$

General relativity predicts values of $\epsilon \approx 0.05{-}0.3$, depending on the spin of the black hole. This efficiency far exceeds that of nuclear fusion and gives rise to the extreme luminosities of quasars, with $L_{\rm bol} \approx 10^{45}{-}10^{49}\,{\rm erg\,s^{-1}}$.

At the same time, the remainder of the accreted mass – i.e., the fraction not radiated away – contributes to the growth of the black hole itself:

$$\dot{M}_{\rm BH} = (1-\epsilon)\dot{M}_{\rm acc}. \tag{1.2}$$

By combining the two above equations, we can directly link the observed luminosity of quasars to the rate at which black holes grow during their active accretion phases, with the only conversion parameter being the radiative efficiency.

What Soltan (1982) recognized is that if one integrates the total light emitted by all quasars over cosmic time, and converts the resulting energy into an accreted mass using a plausible value of ϵ, the result should match the local SMBH mass density inferred from galaxy bulge–black hole scaling relations. For a typical efficiency of $\epsilon \approx 0.1$, the agreement is striking – providing strong evidence that most of the mass in today's SMBHs was assembled through luminous accretion.

This argument leads to a compelling picture in which quasars are direct signposts of black hole growth. The radiation we observe from distant quasars reflects the very process by which SMBHs gain mass. According to this view,

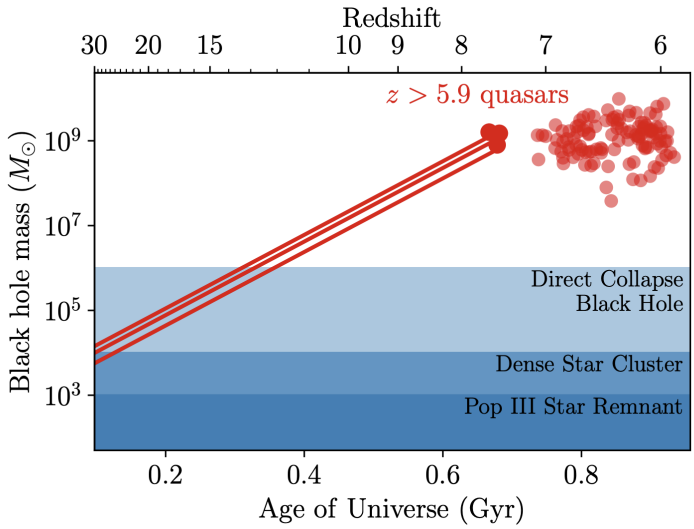

Figure 1.3: Growth histories of SMBHs in the early Universe. Data points indicate black hole masses inferred from observations of quasars at $z > 5.9$. Solid lines show the corresponding growth tracks for the three most distant quasars, assuming continuous, Eddington-limited accretion. Even under this idealized scenario, the observed SMBH masses can only be reached if the initial seed mass $M_{\rm seed}$ is at least $\sim 10^4\,{\rm M}_\odot$ by $z \approx 30$. The shaded blue regions illustrate the typical mass ranges associated with different SMBH seed formation channels. Figure adapted from Fan et al. (2023).

luminous quasars trace the peak phases of SMBH accretion, and their cosmic distribution encodes the history of black hole growth across the Universe. In this way, the Soltan argument establishes a direct connection between the quasar population observed at high redshift and the "relic" SMBHs we find in the centers of galaxies today – as first predicted by Lynden-Bell (1969).

An important consequence of accretion-powered growth is that the radiation emitted by infalling material exerts an outward force – radiation pressure – that counteracts gravity. This interplay naturally sets an upper bound on the accretion rate, beyond which radiation pressure would halt further inflow of gas. This theoretical limit is known as the Eddington limit, and it defines the maximum luminosity an accreting black hole can sustain under the assumption of spherical symmetry and steady inflow.

The Eddington luminosity is derived by equating the outward radiation force on electrons (via Thomson scattering) with the inward gravitational pull on protons:

$$L_{\rm Edd} = \frac{4\pi G M_{\rm BH} m_p c}{\sigma_T} \approx 1.3 \times 10^{38} \left(\frac{M_{\rm BH}}{{\rm M}_\odot}\right) {\rm erg\,s^{-1}}, \tag{1.3}$$

where G is the gravitational constant, m_p is the proton mass, σ_T is the Thomson cross-section, and $M_{\rm BH}$ is the black hole mass. Because $L_{\rm Edd} \propto M_{\rm BH}$, the maximum allowed accretion rate increases linearly with the black hole's mass.

Quasars are observed to radiate at a wide range of Eddington ratios, defined as the ratio of bolometric luminosity to the Eddington luminosity, $\eta = L_{\rm bol}/L_{\rm Edd}$, with the peak of the distribution ranging between $\eta \approx 0.1-1$ depending on the quasars' luminosity and redshift (Wu & Shen 2022). For theoretical modeling, it is common to assume that SMBHs grow at a constant Eddington ratio η, which allows one to derive a characteristic exponential growth law. If a black hole begins with a seed mass $M_{\rm seed}$ at time $t_{\rm seed}$ and accretes continuously at a fixed η, its mass at a later time t is given by:

$$M_{\rm BH}(t) = M_{\rm seed}\, e^{(t-t_{\rm seed})/\tau_{\rm Salp}}, \tag{1.4}$$

where $\tau_{\rm Salp}$ is the Salpeter time, the e-folding timescale for black hole growth under Eddington-limited accretion:

$$\tau_{\rm Salp} \approx 45\,{\rm Myr}\left(\frac{\epsilon}{0.1(1-\epsilon)}\right)\left(\frac{\eta}{1}\right)^{-1}. \tag{1.5}$$

This simple model offers valuable intuition: SMBHs can, in principle, grow from light seed BHs – e.g., $M_{\rm seed} \approx 10^2 - 10^3\,{\rm M}_\odot$, formed from the collapse of the first generation of stars (PopIII stars, Heger et al. 2003) – to billions of solar masses within less than a billion years if accretion proceeds continuously near the Eddington rate.

However, the model also rests on idealized assumptions that likely break down in realistic environments. The complex interplay between accretion and feedback processes makes it unlikely that SMBHs accrete continuously at the same rate. Instead, simulations and physical models show that SMBHs likely grow during discrete, episodic phases (e.g., Novak et al. 2011; Anglés-Alcázar et al. 2015; Trinca et al. 2024). This is often parametrized by introducing an effective duty cycle for quasar activity, which accounts for the fraction of time SMBHs spend in radiatively efficient accretion phases, as opposed to quiescent or inefficient states (Shankar et al. 2009; Pacucci & Loeb 2022). As discussed in Sec. 1.3.2, indirect probes of this intermittent behavior of quasar activity are recently becoming available in the early Universe, opening up the possibility of testing more realistic growth histories against observations (see also Chapter 5).

Additionally, purely accretion-driven models neglect the contribution of black hole mergers. When two SMBHs merge, the resulting remnant has a mass that is approximately equal to the sum of the progenitor masses – reduced by the small fraction of energy radiated away as gravitational waves during the coalescence. While mergers do not alter the integrated black hole mass density that enters the Soltan argument, they can significantly

affect the individual growth histories of SMBHs. This contribution may be particularly important at high redshift and low masses – where merger rates are high – and at low redshift for the most massive black holes whose accretion has been quenched (Volonteri et al. 2003; Pacucci & Loeb 2020; Zou et al. 2024).

1.2.3 The high-redshift frontier

As observational capabilities have pushed the detection of quasars to increasingly earlier cosmic times, a striking realization has emerged: although the number density of quasars declines steeply with redshift (Figure 1.4), some SMBHs with masses $M_{\rm BH} \gtrsim 10^9\,{\rm M}_\odot$ are already in place within the first few hundred million years after the Big Bang. These objects, observed at redshifts $z \gtrsim 6$, rival the most massive black holes found in the centers of present-day galaxies (Fan et al. 2023).

The discovery of ever earlier quasars powered by billion-solar-mass black holes – the current record-holder is at $z \approx 7.64$ (Wang et al. 2021) – has significantly increased the tension with standard models of SMBH formation and growth. The core challenge is straightforward: there appears to be insufficient cosmic time for these black holes to grow from the $\sim 100\,{\rm M}_\odot$ seeds expected from PopIII stellar remnants (Heger et al. 2003) – even under the most optimistic scenario of continuous, Eddington-limited accretion (Haiman & Loeb 2001). This issue is illustrated in Figure 1.3, which compares the expected growth tracks for Eddington-limited accretion to the observed SMBH masses at high redshift.

To resolve this challenge, several massive seed formation scenarios have been proposed (Inayoshi et al. 2020). One leading pathway is the direct collapse of pristine gas clouds into black holes with masses in the range 10^4–10^6 ${\rm M}_\odot$, under specific conditions that suppress fragmentation and prevent star formation (Bromm & Loeb 2003; Volonteri et al. 2008; Latif & Ferrara 2016; Lupi et al. 2021). Another possibility is the runaway collapse of dense stellar clusters, particularly those composed of PopIII stars, where repeated stellar collisions and mergers can lead to the formation of IMBHs (Omukai et al. 2008; Devecchi & Volonteri 2009). These scenarios ease the growth timescale constraints by starting with more massive seeds, but they rely on specific environmental conditions and remain difficult to test observationally.

An alternative route to alleviating the timing problem is to relax the assumption of Eddington-limited accretion. If black holes can grow through super-critical accretion – i.e., with accretion rates higher than the Eddington rate, $\dot{M}_{\rm Edd} = L_{\rm Edd}/\epsilon c^2$ – the stringent time constraints for SMBH growth are considerably relaxed (Volonteri & Rees 2006). Numerical simulations suggest that in dense, gas-rich environments, black holes can exceed the classical Eddington rate under specific physical conditions. Mechanisms

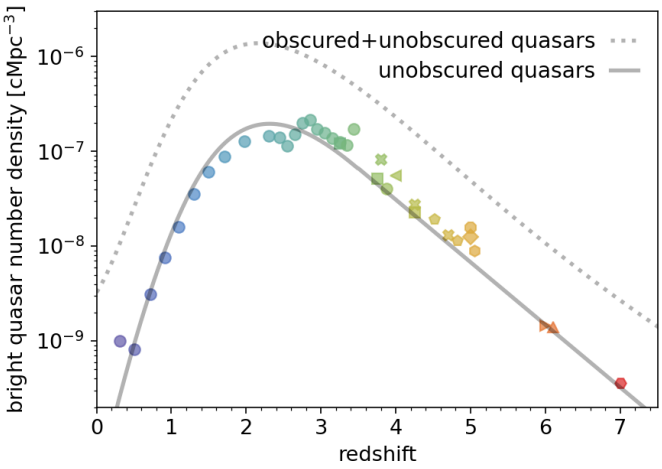

Figure 1.4: Evolution of the number density of bright quasars as a function of redshift. The solid line represents the number density of UV-luminous quasars that are brighter than $M_{1450} < -25.7$, adapted from the global QLF evolution model of Kulkarni et al. (2019) as described in Pizzati et al. (2025). Data points are obtained by integrating UV-optical QLF models at different redshifts. The dotted line shows the number density of bright quasars as estimated from the bolometric QLF model of Shen et al. (2020), by assuming a $L_{\rm bol}$ threshold consistent with the UV magnitude limit mentioned above. The gap between the solid and dotted lines arises from the UV-obscured quasar population. Figure adapted from Schindler et al. (2023) and Pizzati et al. (2025, Chapter 4).

such as photon trapping, slim accretion disks, and anisotropic radiation fields allow accretion to proceed at super-Eddington rates without unbinding the inflowing material (Sądowski et al. 2014; Volonteri et al. 2015; Inayoshi et al. 2020). However, because such accretion is expected to be radiatively inefficient, direct observational confirmation remains challenging. So far, empirical evidence is limited to a few high-redshift quasars with Eddington ratios modestly above unity (Wu et al. 2022). Nonetheless, a growing body of work is investigating indirect signatures of super-critical accretion, either through spectral diagnostics of AGN (Pacucci & Narayan 2024; Lambrides et al. 2024a; Liu et al. 2025; Quadri et al. 2025) or through empirical arguments based on quasar growth timescales and duty cycles (Davies et al. 2019; Eilers et al. 2021, 2024).

Despite substantial theoretical efforts to uncover the growth history of SMBHs, progress remains limited by two persistent challenges: the difficulty of constructing predictive, first-principles models (see Sec. 1.4) and the lack of direct observational constraints beyond the (relatively uncertain) black hole mass estimates. At present, the problem remains highly degenerate – vastly different combinations of seed mass, accretion rate, duty cycle, and

merger history can be fine-tuned to match the observed SMBH masses at a given redshift. As a result, the mere existence of billion-solar-mass black holes in the early Universe is a necessary, but insufficient, condition to distinguish between competing formation and growth scenarios.

Encouragingly, recent and upcoming observations promise to break this impasse and shed new light on SMBH evolution across cosmic time. New AGN candidates uncovered at even higher redshifts by JWST are already putting pressure on existing evolutionary models. Although the nature of these candidates remains uncertain (see Sec. 1.3.4), the advent of wide-field missions such as Euclid and the Roman Space Telescope will soon provide statistically robust samples of luminous quasars at the highest redshifts. At the same time, quasar observables – including luminosity functions, clustering measurements, and proximity zone sizes – are being extended to earlier epochs, offering complementary constraints on black hole accretion physics and environments. Looking further ahead, gravitational wave detections from merging SMBHs will open a fundamentally new observational window into the merger-driven component of black hole growth.

In the following section, I review these recent advancements and highlight the key observational tools – particularly those most relevant to the focus of this thesis – that are currently shaping our understanding of SMBH formation and growth across cosmic time.

1.3 Observations: the evolution of quasars and SMBHs across cosmic times

A major leap forward in quasar studies came with the advent of optical wide-field spectroscopic surveys, which transformed quasars from rare, exotic sources into a population with robust statistical power. Landmark efforts such as the Sloan Digital Sky Survey (SDSS; York et al. 2000), the 2dF QSO Redshift Survey (2QZ; Croom et al. 2004), the Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al. 2013), and its successor, the extended BOSS (eBOSS; Dawson et al. 2016), have collectively catalogued hundreds of thousands of UV-bright quasars across a wide redshift range. These surveys have provided an unprecedented view of the statistical properties and cosmic evolution of quasars and their central SMBHs.

In parallel, multi-wavelength observations have significantly expanded the AGN census beyond optically selected quasars. Owing to their broadband emission, AGN can be detected across the entire electromagnetic spectrum, and surveys at other wavelengths – X-ray missions (e.g., Chandra, XMM-Newton), radio surveys (e.g., FIRST, NVSS, LOFAR), and mid-infrared campaigns (e.g., Spitzer, WISE) – have uncovered substantial populations of AGN that are obscured in the UV-optical. These datasets have been

instrumental in tracing black hole accretion across a wide range of host galaxy environments and evolutionary stages. Together, they provide a critical foundation for building a more complete and less biased picture of SMBH growth across cosmic history (Padovani et al. 2017).

1.3.1 The quasar luminosity function

The most basic metric for characterizing the demographic properties of the quasar population is the quasar luminosity function (QLF). The QLF describes the comoving number density of quasars as a function of luminosity and has been extensively measured across multiple wavelengths and over a wide range of redshifts (e.g., Boyle et al. 2000; Richards et al. 2006; Ross et al. 2013; Akiyama et al. 2018).

Empirically, the QLF is typically modeled as a broken power law, with the normalization, faint-end slope, and characteristic break luminosity evolving strongly with redshift. This evolution reflects the cosmic history of black hole accretion, with a pronounced peak at $z \sim 2-3$ – the so-called "cosmic noon" – when both quasar activity and global star formation reach their maximum (Shen et al. 2020). The striking similarity between the redshift evolution of the QLF and that of the cosmic star formation rate density (Ciotti et al. 2003) supports the widely discussed scenario of co-evolution between SMBHs and their host galaxies (Merloni & Heinz 2008; Heckman & Best 2014). Figure 1.4 illustrates the evolution of the bright quasar population across cosmic time, highlighting its rise from the epoch of reionization, peak activity around cosmic noon, and subsequent decline.

The most robust constraints on the QLF come from optical and UV-selected quasar samples, where the luminosity function is often expressed in terms of absolute UV magnitude (e.g., M_{1450}; Kulkarni et al. 2019). However, these measurements are inherently biased toward unobscured quasars, as obscuration from dust and gas can significantly attenuate emission in the UV and optical bands. This bias complicates the interpretation of QLF measurements, particularly because the fraction of obscured AGN is observed to vary with both luminosity and redshift (Aird et al. 2015; Buchner et al. 2015). These effects introduce systematic uncertainties when comparing observations with theoretical models, which typically predict intrinsic, bolometric luminosities.

To overcome these limitations, recent efforts have focused on constructing multi-wavelength AGN samples that combine X-ray, mid-infrared, UV-optical, and radio observations. These datasets are essential for recovering obscured quasars that are systematically missed in UV-only surveys. In particular, X-ray data allow for population-level obscuration corrections through measurements of hydrogen column densities (Ueda et al. 2014), enabling estimates of intrinsic AGN luminosities even for heavily absorbed systems. By applying these corrections and synthesizing observations across

multiple bands, several studies have reconstructed the bolometric QLF, providing a more comprehensive benchmark for models of black hole growth (Hopkins et al. 2007a; Shen et al. 2020). In Figure 1.4, we compare a UV-selected QLF model (solid line) with a bolometric QLF derived from multi-wavelength data (dashed line). While these bolometric reconstructions offer a substantially improved census of AGN activity, significant uncertainties persist – especially at high redshift, where obscuration properties remain poorly constrained and the number of detected sources is still limited.

1.3.2 Quasar clustering and the duty cycle of quasars

In addition to the QLF, another key statistical observable for characterizing quasars in a cosmological context is their large-scale clustering. As discussed in Section 1.1.1, the ΛCDM paradigm predicts a strong dependence of halo clustering on mass. This implies that measuring the clustering strength of a population provides a powerful way to infer the characteristic mass of its host dark matter halos. By comparing the clustering amplitude of quasars to that of halos across a range of masses, one can constrain the typical environments in which quasars reside. In Chapters 2 and 3, we will build on this approach by jointly modeling the QLF and quasar clustering to infer the full mass distribution of quasar host halos – a quantity we refer to as the quasar–host mass function (QHMF).

The advent of large spectroscopic quasar surveys has enabled precise measurements of the quasar two-point auto-correlation function, the most direct probe of quasar clustering on cosmological scales. Numerous studies have characterized this clustering across a wide range of redshifts (e.g., Porciani et al. 2004; Croom et al. 2005; Porciani & Norberg 2006; Shen et al. 2007; da Ângela et al. 2008; Ross et al. 2009; White et al. 2012; Eftekharzadeh et al. 2015), consistently finding that quasars typically inhabit dark matter halos with masses around 10^{12}–$10^{13}\,\mathrm{M}_{\odot}$. This characteristic halo mass appears to be largely independent of quasar luminosity and evolves only mildly with redshift (see Figure 1.5). A possible exception to this is the large host halo mass inferred by Shen et al. (2007) at $z \approx 4$, which suggests a rapid evolution of the quasar properties at high redshift and provides very tight constraints on the inferred host mass distribution (Pizzati et al. 2024a). Chapters 2 and 3 discuss in detail the implications of this measurement, building up on previous work from White et al. (2008); Wyithe & Loeb (2009); Shankar et al. (2010b).

Interestingly, quasar clustering measurements not only constrain the typical mass of quasar host halos but also offer insights into the integrated timescale of quasar activity by estimating the quasar *duty cycle* (Martini & Weinberg 2001; Haiman & Hui 2001). The concept is illustrated in Figure 1.6: if SMBH accretion proceeds as a stochastic process, then the observed quasar population at any given epoch represents a random, luminous subset of the

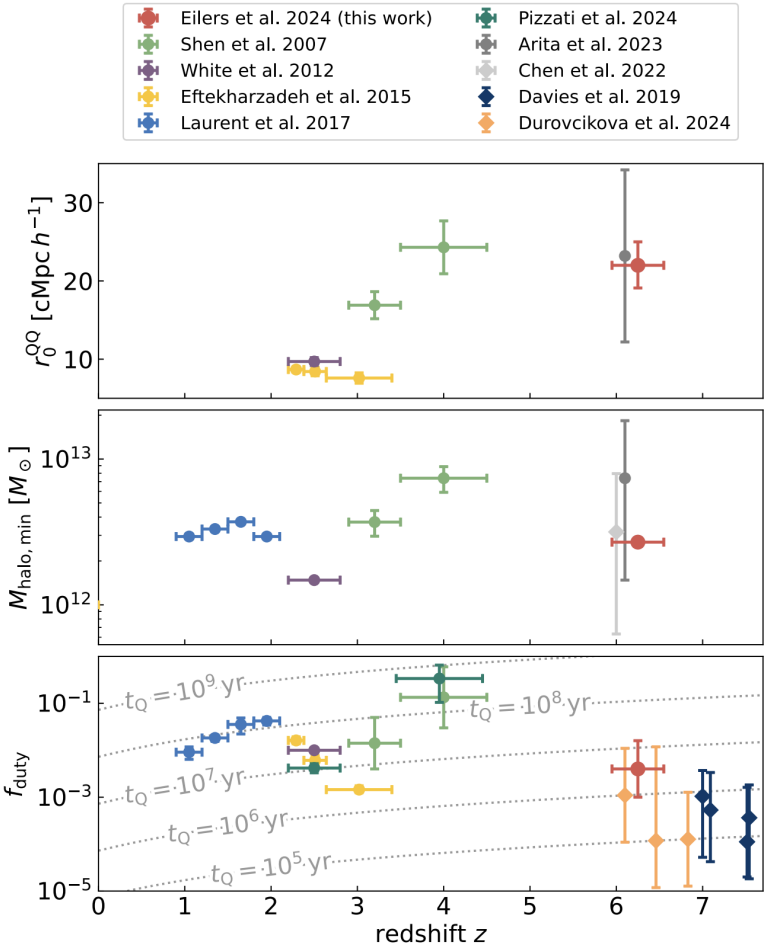

Figure 1.5: Redshift evolution of three key quasar properties: the autocorrelation length (top), host dark matter halo mass (middle), and duty cycle (bottom), as a function of redshift. Measurements at $z \approx 6$ are all based on recent work and obtained using different methods. The halo mass estimate from Chen et al. (2022) is based on transmitted flux measurements along quasar sightlines, and the duty cycle measurements from Davies et al. (2019); Ďurovčíková et al. (2024) are obtained from Lyα damping wing analyses. The remaining data are based on clustering measurements: Arita et al. (2023) measured the auto-correlation function of faint $z \approx 6$ quasars, while Eilers et al. (2024) used JWST WFSS observations to estimate the quasar-galaxy cross-correlation function at the same redshift. Gray dotted lines in the bottom panel correspond to constant quasar lifetimes. The results show that the characteristic host halo mass of quasars sits between $\sim 10^{12}\,\mathrm{M}_\odot$ and $\sim 10^{13}\,\mathrm{M}_\odot$ at all redshifts. The corresponding duty cycle, however, shows a significant redshift evolution. At $z \gtrsim 6$, f_{duty} is found to be $\lesssim 1\%$ using independent methods. Figure taken from Eilers et al. (2024).

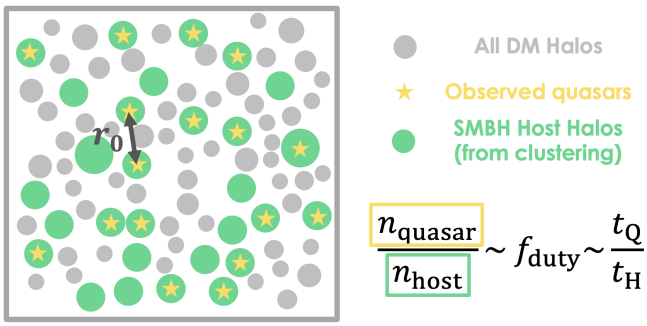

Figure 1.6: Schematic illustration of how quasar clustering measurements can be used to infer the quasar duty cycle. The full population of dark matter halos is shown as gray circles. Quasars (yellow stars) occupy only a subset of these halos, representing a stochastic sampling of the underlying SMBH population. By measuring the clustering of quasars, one can infer the typical mass – and hence number density – of their host halos (highlighted in green). Comparing the observed number density of quasars ($n_{\rm quasar}$) to the number density of similarly clustered halos ($n_{\rm hosts}$) yields an estimate of the quasar duty cycle: the average fraction of time a SMBH is observed as an active quasar. This duty cycle is commonly expressed as the ratio between the integrated quasar lifetime, $t_{\rm Q}$, and the cosmic timescale over which halos exist, approximated by the Hubble time, $t_{\rm H}$.

full SMBH population. Subsampling a population does not change its clustering: hence, by measuring the clustering of quasars we can infer the clustering of the dark matter halos hosting SMBHs. Using the connection between clustering and halo mass, we can then infer the characteristic mass and number density of these SMBH-host halos. We are implicitly assuming here that all massive halos host a SMBH (which can be either active or dormant) at their center, as it is the case in the local Universe. Comparing the number density of SMBH hosts ($n_{\rm host}$) to the observed number density of quasars ($n_{\rm quasar}$) yields the fraction of SMBHs that are active (i.e., visible as quasars) at a given time. This fraction corresponds to the average time a black hole spends in an active, luminous quasar phase over cosmic time: the so-called quasar duty cycle ($f_{\rm duty}$).

Recent efforts have extended clustering and duty cycle measurements into the epoch of reionization, a long-sought goal given the importance of understanding the environments of early quasars. While many studies have attempted to assess whether high-redshift quasars reside in overdense regions (e.g., Kim et al. 2009; Simpson et al. 2014), the first robust measurements of quasar clustering at $z \gtrsim 6$ have only recently emerged. These include the quasar auto-correlation analysis by Arita et al. (2023), based on faint quasars from the SHELLQs survey, and the quasar–galaxy cross-correlation study by Eilers et al. (2024), enabled by JWST's Wide Field Slitless Spectroscopy (WFSS) mode. The primary focus of Chapter 3 is to model these

measurements[2] and study their implications for quasar activity and SMBH growth in the early Universe. Intriguingly, the inferred duty cycle at $z \sim 6$ is $f_{\rm duty} \lesssim 1\%$, in agreement with independent estimates of quasar activity timescales from Lyα damping wing and proximity zone analyses (Davies et al. 2019; Ďurovčíková et al. 2024). This poses a challenge to standard scenarios of continuous, Eddington-limited black hole growth that are often invoked to explain the rapid emergence of $\gtrsim 10^9\,{\rm M}_\odot$ SMBHs in the early Universe.

1.3.3 SMBH mergers and gravitational waves

While electromagnetic observations of quasars trace the accretion-driven growth of SMBHs, gravitational wave (GW) observatories offer a complementary probe by directly accessing the merger-driven channel of black hole growth. As discussed in Sec. 1.2.1, ground-based GW detectors such as LIGO and Virgo have already provided a wealth of information about the stellar-mass black hole population through detections of compact binary coalescences (Abbott et al. 2023). In contrast, progress in the regime of SMBHs is more recent, but has accelerated markedly with the advent of pulsar timing array (PTA) experiments and will mature substantially in the next decade.

A landmark step forward has been the recent detection of a stochastic gravitational wave background (GWB) by multiple PTA collaborations, including NANOGrav (Agazie et al. 2023), the European PTA (EPTA Collaboration et al. 2023), the Parkes PTA (Reardon et al. 2023), and the Chinese PTA (Xu et al. 2023). These detections reveal a common-spectrum red noise process in pulsar timing residuals with evidence for spatial correlations consistent with the Hellings & Downs (1983) signature expected from gravitational waves in general relativity (Burke-Spolaor et al. 2019). The detected GWB is broadly consistent with the combined emission from inspiraling supermassive black hole binaries, typically with masses of order $\sim 10^8$–$10^{10}\,{\rm M}_\odot$ at redshifts $z \lesssim 2$. This detection provides the first direct GW evidence for the prevalence of SMBH mergers in the nanohertz frequency regime, marking a pivotal step for the field of SMBH evolution.

Looking ahead, PTAs are expected to move beyond statistical detections of a stochastic background to the resolution of individual SMBH binary systems. These future detections will be most sensitive to the most massive binaries ($M_{\rm tot} \gtrsim 10^9\,{\rm M}_\odot$) at relatively low redshifts ($z \lesssim 1$), and could enable the measurement of dynamical properties such as orbital eccentricity, mass ratios, and the nature of the merger environment. Concurrently, improvements in PTA sensitivity will enhance constraints on the unresolved

[2] In Appendix D of Chapter 3, I demonstrate that the halo mass and duty cycle estimates reported by Arita et al. (2023) are subject to methodological issues and are therefore unreliable.

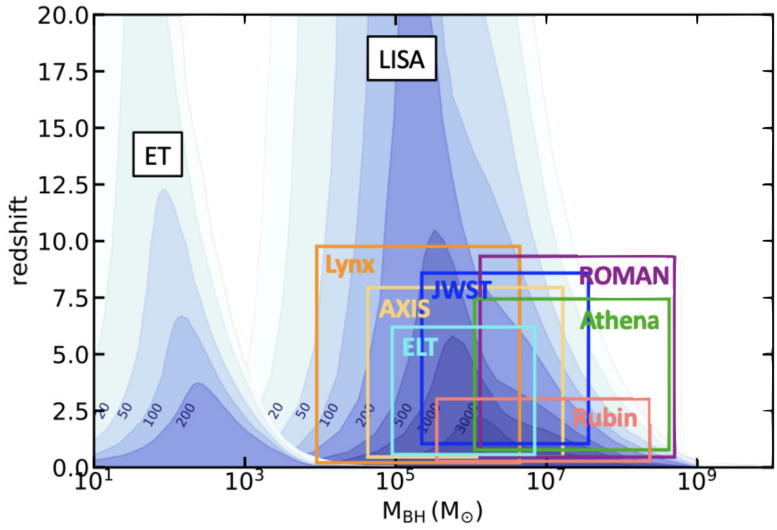

Figure 1.7: A broad overview of the black hole mass–redshift landscape and the observational capabilities expected to probe it. The figure shows the sensitivity ranges (i.e., the expected signal-to-noise curves for non-spinning binaries with mass ratio 0.5) of various gravitational wave (GW) observatories, such as LISA and future third-generation ground-based interferometers like the Einstein Telescope (ET), for detecting black hole mergers across a wide range of masses ($\sim 10^1$–$10^7\,M_\odot$) and cosmic epochs. These GW detectors will enable full-sky surveys, reaching from the local Universe to the era of the first black holes. Overlaid are the approximate reach of current and planned electromagnetic (EM) facilities – such as JWST, Roman, Rubin, the ELT, and next-generation X-ray observatories like Athena, LynX, and AXIS – highlighting the synergy between EM and GW observations. Taken from Volonteri et al. (2021).

gravitational wave background, offering a critical test of theoretical models of SMBH evolution. Intriguingly, several recent studies have reported emerging discrepancies between the amplitude of the detected background and predictions from leading semi-analytic and hydrodynamic models of SMBH assembly (e.g., Lapi et al. 2025). These tensions may reflect previously unmodeled physical processes, such as stalling of SMBH binaries due to inefficient hardening, coupling with circumbinary gas disks, or inaccuracies in merger rate prescriptions. As PTA datasets continue to grow in precision and temporal coverage, they will provide an increasingly powerful probe of the physics driving SMBH mergers and their role in cosmic structure formation.

At the opposite end of the SMBH mass spectrum lies the future space-based mission LISA (Laser Interferometer Space Antenna). Unlike PTAs, which are sensitive to the mergers of the most massive SMBHs at low redshift, LISA will probe the coalescence of lower-mass SMBHs and IMBHs

in the range 10^4–$10^7\,M_\odot$, reaching out to redshifts as high as $z \gtrsim 10$ with exceptional signal-to-noise. This makes LISA uniquely suited to explore the early formation and assembly of SMBHs, as well as the long-sought IMBH population. In particular, LISA will offer key insights into seed formation channels, early merger rates, and black hole occupation fractions in low-mass galaxies – regimes currently inaccessible to electromagnetic or PTA observations. Numerous studies have forecasted LISA's potential to constrain black hole demographics and binary environments (e.g., Sesana et al. 2007; Tanaka & Haiman 2009; Amaro-Seoane et al. 2023; Wang et al. 2025), but the field remains highly uncertain, with predicted detection rates spanning several orders of magnitude.

Figure 1.7 offers a schematic view of the observational landscape, mapping the redshift and black hole mass ranges accessible to current and upcoming gravitational wave and electromagnetic observatories. Alongside LISA, it includes the Einstein Telescope (ET) as a representative of the planned third generation (3G) of ground-based GW interferometers. These instruments will extend sensitivity to stellar-mass/IMBH mergers at high redshift, complementing LISA's reach and contributing to a unified picture of black hole growth across cosmic history. The design and associated challenges of 3G detectors are discussed in more detail in Chapter 6.

1.3.4 New challenges in the JWST era: the nature of "little red dots" and other broad-line AGN

The launch of the James Webb Space Telescope (JWST) has ushered in a transformative era for the study of AGN and quasars at high redshift. Thanks to its unprecedented sensitivity in the infrared, JWST can detect AGN that were previously invisible to traditional rest-frame UV and optical surveys, which until now have dominated our view of the AGN population during the epoch of reionization (Fan et al. 2023).

The first few years of observations with JWST have indeed uncovered a surprisingly rich population of faint AGN candidates at $z \approx 4$–10 (e.g., Harikane et al. 2023; Maiolino et al. 2024; Übler et al. 2023; Kocevski et al. 2023; Matthee et al. 2024b; Greene et al. 2024; Bogdán et al. 2024). Many of these sources are identified via the presence of broad rest-frame optical emission lines such as Hα or Hβ, accessible for the first time at high redshift thanks to JWST's NIRSpec and NIRCam instruments. These features enable black hole mass estimates of $M_{\rm BH} \gtrsim 10^{6-8}\,M_\odot$ and bolometric luminosities of $L_{\rm bol} \gtrsim 10^{43-46}\,{\rm erg\,s^{-1}}$, extending the census of actively accreting SMBHs well below the luminosities of previously known quasars at comparable redshifts. The emergence of this population opens new avenues to address key questions in early black hole and galaxy evolution, including the co-evolution of SMBHs and their hosts (e.g., Pacucci et al. 2023), the contribution of faint AGN to hydrogen reionization (e.g., Dayal et al. 2024; Madau et al.

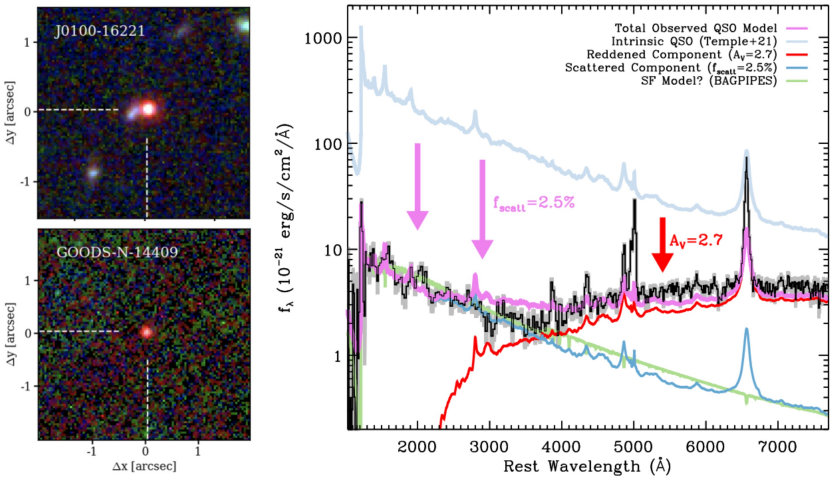

Figure 1.8: *Left:* False-color NIRCam images of two representative "little red dots" (LRDs) from the broad-line AGN sample of Matthee et al. (2024b), showcasing their compact, red morphology. *Right:* Prism/NIRSpec spectrum (black line) of a representative LRD (object MSAID4286), exhibiting the characteristic "V-shaped" spectral energy distribution. The steep red continuum is explained by a heavily reddened AGN template (red line), while the rising UV slope cannot be accounted for by the intrinsic AGN continuum alone. Two alternative components are shown to model the UV excess: (i) scattered AGN light at the ∼2.5% level (light blue), and (ii) moderate star formation from a stellar population model. Figure taken from Greene et al. (2024).

2024), and the underlying seeding and growth pathways of SMBHs in the early Universe (e.g., Li et al. 2024).

A particularly intriguing subset of the JWST AGN candidates – estimated to comprise $\gtrsim 20\%$ of the sample (Harikane et al. 2023; Taylor et al. 2024) – exhibits unusually steep, red continua in the rest-frame optical, along with compact morphologies. These objects have become known as "little red dots" (LRDs). Figure 1.8 presents two representative NIRCam images of such sources from Matthee et al. (2024b), as well as a PRISM/NIRSpec spectrum of an LRD from Greene et al. (2024). The latter illustrates the characteristic "V-shaped" spectral energy distribution (SED), which is created by the combination of a blue component in the rest-frame UV and the rising red continuum in the rest-frame optical. In the pioneering work of Greene et al. (2024), this is interpreted as a combination of a reddened quasar template – accounting for the steeply rising red continuum – and an additional blue component, attributed either to starlight from the host galaxy or to quasar light scattered into our line of sight.

Interestingly, this and other interpretations suggest moderate levels of dust attenuation, typically in the range $A_V \approx 1{-}4$ (Kokorev et al. 2024a; Greene et al. 2024). When the SEDs and emission lines of LRDs are corrected

for this extinction, the inferred bolometric luminosities and black hole masses are found to be comparable to those of UV-selected, unobscured quasars identified in pre-JWST surveys (e.g., Fan et al. 2023; Matsuoka et al. 2022). This similarity is striking given the vastly different selection strategies and survey volumes. While UV-luminous quasars at $z \gtrsim 6$ have been discovered in wide-field surveys covering $\sim$ 1400 deg^2 (Matsuoka et al. 2022), the JWST-detected LRDs are emerging from pencil-beam or small fields totaling only $\sim$ 300–600 arcmin^2 (e.g., Matthee et al. 2024b; Kokorev et al. 2024a). This implies an apparent overabundance of LRDs by factors of 10^3–10^4, assuming the populations are otherwise comparable – a discrepancy that cannot be easily explained.

This dramatic difference in inferred number densities raises fundamental questions about the nature of LRDs and JWST-selected AGN candidates more broadly. The possibility that they trace a large population of obscured, broad-line quasars that was previously undetected is challenged further by the unusual SED properties of LRDs beyond the rest-frame UV-optical. Multiple studies have reported a set of anomalous features that distinguish LRDs and other JWST AGN candidates from classical quasars, including apparent X-ray weakness (e.g., Yue et al. 2024b), the abundant presence of Balmer absorption features as well as strong Balmer breaks (e.g., Juodžbalis et al. 2024; de Graaff et al. 2025), a potential lack of variability (Kokubo & Harikane 2024), and unexpectedly faint dust emission in both the mid- and far-infrared (e.g., Pérez-González et al. 2024; Casey et al. 2025).

The enigmatic nature of these sources has sparked widespread interest within the community, leading to a steady stream of papers proposing a broad spectrum of theoretical interpretations for the nature of LRDs – ranging from exotic scenarios to more conservative models invoking complex geometries and radiative transfer effects. A simple NASA/ADS search reveals that over $\sim$150 preprints and peer-reviewed articles in the last two years mention "little red dots" in their title or abstract, underscoring both the scale of engagement and the unsettled nature of this emerging field. While this is not the place to review the full breadth of this rapidly growing literature, the sheer diversity of proposed explanations highlights just how far we remain from a definitive understanding.

Nevertheless, some common themes are beginning to crystallize. Several recent studies suggest that the Balmer absorption features, pronounced Balmer break, and – in some cases – the steep rest-frame optical continua observed in LRDs may arise from dense, cool gas enshrouding a rapidly accreting SMBH (Inayoshi & Maiolino 2025; Naidu et al. 2025) – potentially indicative of super-Eddington accretion flows (e.g., Liu et al. 2025) or even SMBH seeding in the early Universe (e.g., Begelman & Dexter 2025). On the other hand, some studies have questioned whether LRDs host SMBHs at all. In these cases, the broad emission lines – typically seen as signatures of AGN activity – could instead be powered by compact gas in extreme

starburst or rare transient phenomena unrelated to SMBH accretion (e.g., Baggen et al. 2024; Sacchi & Bogdan 2025).

Chapter 4 of this thesis is dedicated to investigating LRDs. Rather than attempting to define what they are, I take a complementary approach – arguing what *they are not*. Using an argument that combines quasar clustering with the observed number density of LRDs, I demonstrate that LRDs cannot merely be obscured counterparts of UV-luminous quasars. The reason is simple: LRDs are too abundant to reside in the same halos where UV-luminous quasars live. Thus, they need to follow intrinsically different scaling relations between SMBHs, host galaxies, and halos than those established for quasars. This conclusion is now supported by emerging clustering measurements, which show that LRDs exhibit spatial correlations consistent with typical field galaxies, in stark contrast to the strong clustering seen in quasars at similar redshifts (Arita et al. 2025; Matthee et al. 2024a; Lin et al. 2025). These results strongly suggest that LRDs are not just dust-obscured versions of the quasars we already know – but instead represent a fundamentally distinct population. They may trace a different evolutionary pathway in SMBH and galaxy growth, or reflect a phase of black hole formation or evolution driven by fundamentally different physical mechanisms. Perhaps they are not accreting SMBHs at all, and could instead reveal something new about extreme stellar processes. Uncovering their true nature remains an exciting and urgent challenge for the field.

1.4 Theoretical models: key uncertainties and future directions

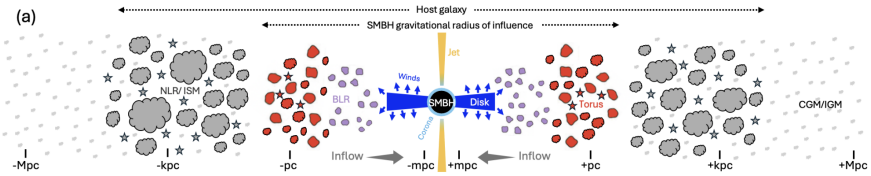

Figure 1.9: Schematic illustration of an AGN (vertical axis is not to scale) embedded within its larger-scale host galaxy and dark matter halo environment, spanning (logarithmic) scales from milli-parsecs to mega-parsecs along the horizontal axis. Key components are marked, including the regions dominated by the SMBH's gravitational influence – the AGN itself: accretion disk, X-ray "corona," possible winds or jets from the disk, broad-line region (BLR), and dusty molecular "torus" – as well as the surrounding host galaxy and halo. Figure adapted from Alexander et al. (2025).

The biggest challenge faced by any first-principles model of SMBH growth in a cosmological context is that of scale. This is illustrated in Figure 1.9, which sketches the vast range of spatial scales involved in SMBH

evolution. Accretion onto SMBHs occurs in a disk that extends from just a few Schwarzschild radii (R_S[3]) out to a few hundred R_S. The AGN central engine also comprises a compact broad-line region (BLR) and a dusty molecular torus – together rarely extending beyond ~ 1 pc. Even so, most of the intrinsic UV-optical/X-ray emission in a type I AGN is confined within the innermost ~ 0.01 pc, a region not much larger than our Solar System.

Although accretion is regulated by processes on the smallest scales, the long-term growth of SMBHs is ultimately driven by galactic and cosmological mechanisms operating across kpc to tens-of-Mpc scales. The enormous energy output from accretion further couples the central AGN regions to the large-scale environment: AGN radiation and feedback can impact structures that are several Mpc away. Reproducing the formation and evolution of SMBHs and quasars in a cosmological framework would thus require resolving both the minuscule scales of accretion and the vast cosmic volumes where such rare objects statistically arise – a task that spans twelve orders of magnitude and will remain computationally unfeasible in the foreseeable future. Even the most advanced zoom-in cosmological hydrodynamical simulations, which focus computational resources on selected halos, typically achieve resolutions no better than ~ 1 pc in the nuclear region[4], and must therefore rely on sub-grid prescriptions to model the unresolved physics of AGN fueling and feedback (Anglés-Alcázar et al. 2021).

Given the extreme range of spatial and physical scales involved, it is perhaps unsurprising that, despite major progress over the past decade in reproducing global galaxy populations and their evolution (see Vogelsberger et al. 2020, for a review), hydrodynamical simulations still struggle to match observational constraints on SMBHs. Most large-scale simulations are calibrated to reproduce local SMBH–galaxy scaling relations – and generally do so with reasonable accuracy (e.g., Di Matteo et al. 2005; Booth & Schaye 2009) – but they diverge widely in their predictions for how SMBHs grow and evolve over cosmic time (Habouzit et al. 2021, 2022; Porras-Valverde et al. 2025). While the sub-grid treatments of star formation – operating on scales of giant molecular clouds, i.e., $\sim 10-100$ pc – are relatively well studied and established, physical models for SMBH growth and AGN feedback remain much more rudimentary and vary substantially across different simulation frameworks.

In recent years, the simulation community has invested considerable effort in refining the treatment of AGN within cosmological simulations, aiming to better connect the physics of SMBH accretion with the broader processes

[3]The Schwarzschild radius of a black hole is given by $R_S = 2M_{BH}/c^2$.

[4]Recent advances now make it possible to resolve the accretion disc of a SMBH within a cosmological simulation by recursively refining the central region (see, e.g., Hopkins et al. 2025). Such simulations, however, can only follow the system over very short timescales, in contrast to standard zoom-in simulations that capture the full cosmological evolution of selected halos.

of galaxy formation and evolution. Notable progress has been made, with important developments proposed in the modeling of black hole seeding (e.g., Bhowmick et al. 2024), dynamics (e.g., Genina et al. 2024), accretion (e.g., Koudmani et al. 2024; Weinberger et al. 2025), and feedback (e.g., Huško et al. 2024). However, despite these advances, significant uncertainties remain, and the predictive power of current simulations for SMBH evolution remains limited.

An alternative to modeling SMBH evolution through first-principles physics is the use of phenomenological or empirical models. These models are agnostic to the detailed physical mechanisms driving black hole growth, instead aiming to capture the observed evolution of SMBHs through simple, parametric formulations that are constrained directly by data. By relying on observations, these models seek to empirically characterize the demographics of SMBHs and their connection to host galaxies in a self-consistent way. While they have limited predictive power beyond the range of the observational constraints they are built upon, they are powerful tools for identifying the key empirical trends that any physical model must reproduce.

Empirical models involving SMBHs can be broadly divided into two main categories. The first class focuses on the evolution of the SMBH mass function by solving the continuity equation. These models rely on two key ingredients: the local observed SMBH masses and the redshift-dependent QLFs. They extend the Soltan (1982) argument to reconstruct the evolution of the SMBH population by linking their growth history to the observed energy output, yielding estimates of key quantities such as the average radiative efficiency, duty cycle, and the distribution of Eddington ratios. Seminal work in this area includes studies by Yu & Tremaine (2002); Merloni & Heinz (2008); Shankar et al. (2009), as well as the more recent models by, e.g., Aversa et al. (2015); Tucci & Volonteri (2017). Collectively, these studies helped to build a statistical picture of SMBH growth across cosmic time.

The second class of models centers on the co-evolution of galaxies and SMBHs, leveraging the empirical relations between galaxy and black hole properties. These models often build upon semi-empirical frameworks originally developed for galaxy evolution, extending them to incorporate SMBHs and AGN activity by assuming a (redshift-dependent) SMBH–galaxy relation. Classic implementations of this approach include Croton (2009), Conroy & White (2013), and Caplar et al. (2015). A notable recent example is TRINITY (Zhang et al. 2023b), which extends the halo-galaxy connection formalism of Behroozi et al. (2013) to simultaneously model the evolution of dark matter halos, galaxies, and SMBHs. TRINITY assumes that SMBH growth follows a redshift-evolving relation with galaxy properties and constrains average black hole evolution histories by requiring consistency with a broad range of SMBH and galaxy observables.

Phenomenological models are inherently constrained by the quality and completeness of the observational data they are built upon. In this sense, they reflect the current limits of our ability to probe SMBH evolution. Yet this dependence on data also offers a key strength: as observations improve – particularly at high redshift – these models can be continually refined, allowing for more accurate and insightful reconstructions of black hole growth across cosmic time.

Motivated by the growing body of current and upcoming observations targeting quasars and SMBHs in the early Universe, in Chapter 5 I develop an empirical modeling framework designed to flexibly and self-consistently interpret a wide range of quasar observables. These include luminosity functions, Eddington ratio distributions, and large-scale clustering – including new measurements at high redshift (Sec. 1.3) – all integrated within a unified model built upon a large-volume, dark-matter-only cosmological simulation. By reproducing the diversity of individual black hole growth histories and quasar light curves, this approach enables a systematic exploration of the physical mechanisms that drive SMBH evolution and sheds light on the key processes shaping the high-redshift quasar population.

1.5 This thesis

This thesis presents six studies conducted in collaboration with my co-authors, covering a broad range of topics: the clustering and evolution of high-redshift quasars and supermassive black holes (SMBHs), the emerging population of "little red dots" and JWST-selected AGN candidates, the study of parameter inference for third-generation gravitational wave detectors, and the morphology of protoplanetary discs. While most chapters have already been introduced in earlier sections, I provide below a summary of each chapter – highlighting its objectives and main results – to serve as a reference for navigating the structure of this thesis.

Chapter 2 presents a framework we developed to jointly model the clustering and luminosity function of quasars at arbitrary redshifts. This framework is built on large-volume, dark-matter-only cosmological simulations, and includes a method to extract the relevant halo statistics – i.e., the halo mass function and the cross-correlation of halos with different masses – while combining multiple simulations to increase the dynamic range. With this method, we can effectively probe the largest scales which are needed to find rare, massive halos while retaining the capability of modeling less massive and more common systems.

The primary goal of this chapter is to revisit the quasar clustering measurements reported by Shen et al. (2007), which revealed an unusually strong clustering signal at $z \approx 4$. We demonstrate that reproducing both the quasar luminosity function (QLF) and the clustering amplitude at this

redshift is possible, but only under the extreme assumption that nearly all quasars occupy the most massive dark matter halos. While modeling the QLF or clustering separately admits a wide range of parameter choices, jointly fitting the two observables substantially tightens the constraints – pointing to a quasar duty cycle close to unity and a remarkably small scatter in the luminosity-halo mass relation. These conclusions are both striking and challenging, underscoring the need for new observational campaigns to either confirm or refute the Shen et al. (2007) results. Finally, we highlight the flexibility of our framework by applying it to the quasar clustering measurements at $z \approx 2.5$ from Eftekharzadeh et al. (2015), which provide some of the most precise constraints on quasar clustering currently available.

In **Chapter 3**, we extend the model introduced in Chapter 2 to incorporate the population of line-emitting galaxies observed in JWST WFSS/NIRCam surveys. In particular, we focus on [O III] emitters identified in JWST surveys such as EIGER (Kashino et al. 2023). Capturing both quasars and galaxies within a unified framework requires an exceptionally large cosmological volume, which we achieve using the FLAMINGO-10k dark-matter-only simulation (Schaller et al., in prep.), specifically designed for this purpose. Leveraging this simulation, we successfully reproduce several key observables at $z \approx 6$: the luminosity functions of quasars and [O III] emitters (Schindler et al. 2023; Matthee et al. 2023), their respective auto-correlation functions (Arita et al. 2023; Eilers et al. 2024), and the quasar–galaxy cross-correlation function (Eilers et al. 2024).

The model yields predictions for the luminosity–halo mass relation, host halo mass distributions, and duty cycles for both quasars and [O III] emitters. To our knowledge, this is the first study to constrain the properties of these populations at such high redshifts using clustering measurements. Notably, our results point to a very low quasar duty cycle at $z \approx 6$ ($f_{\rm duty} \lesssim 1\%$). We discuss the implications of these findings for early SMBH and galaxy formation, and highlight the puzzling evolution of quasar properties implied by measurements of quasars clustering at different cosmic epochs.

Chapter 4 investigates a newly emerging population of broad-line AGN candidates identified in deep JWST imaging and spectroscopy – some of which exhibit a steep rest-frame optical continuum and are thus referred to as "little red dots" (LRDs; Matthee et al. 2024b). After correcting for obscuration, many LRDs exhibit bolometric luminosities comparable to UV-selected quasars, despite being detected in surveys covering areas thousands of times smaller (Greene et al. 2024). This striking contrast implies that LRDs are significantly more abundant than unobscured quasars of similar luminosity, posing a major challenge to existing models of SMBH growth and AGN activity in the early Universe (Inayoshi & Ichikawa 2024).

Through a detailed comparison between JWST-selected AGN/LRDs and UV-selected quasars, we conclude that LRDs outnumber quasars by a large and rapidly evolving factor with redshift. Interestingly, this suggests

that the large population of LRDs cannot be accommodated in the same halos where unobscured quasars live, suggesting that LRDs represent a distinct evolutionary phase of SMBH growth, governed by different black hole–galaxy–halo scaling relations. Supporting this interpretation, recent clustering measurements show that LRDs exhibit spatial correlations consistent with typical star-forming galaxies, in stark contrast to the strong clustering seen in quasars at similar redshifts (Arita et al. 2025; Matthee et al. 2024a; Lin et al. 2025). Together, these findings indicate that LRDs are not merely obscured versions of known quasars, but instead constitute a fundamentally distinct population in the early AGN landscape – or possibly, that they are not AGN at all.

Chapter 5 builds on the models developed in Chapters 2 and 3 by introducing an evolutionary framework for SMBHs and quasars embedded within a large dark-matter-only cosmological simulation. As in previous chapters, the model is constrained by key quasar observables – namely, the luminosity function and large-scale clustering – but is now applied consistently across all redshifts within a unified framework. Additionally, SMBH mass measurements (or equivalently, the Eddington ratio distribution function) are incorporated as an independent constraint, in a way that mitigates biases due to the limited completeness of current observations. At its core, the model connects the growth history of each SMBH to that of its host halo through parametric functions that account for both average evolutionary trends and stochastic variability. SMBH growth is treated self-consistently, with accretion directly driving quasar activity.

Despite its simplicity, the model successfully reproduces a broad range of observables from the epoch of reionization ($z \approx 7$) down to cosmic noon ($z \approx 2$). We focus in particular on the early buildup of the most massive SMBHs ($\gtrsim 10^9\,\mathrm{M}_\odot$ by $z \approx 7$), and investigate the primary drivers of this growth – including the relative contributions of accretion and mergers, as well as the role of the accretion coherence timescale. Future extensions of this framework will target lower redshifts and incorporate additional observational constraints, such as quasar proximity zones and damping wing measurements, and the gravitational wave background detected by pulsar timing array (PTA) experiments.

Chapter 6 focuses on parameter inference for gravitational wave (GW) signals in the era of third-generation detectors, such as the Cosmic Explorer (CE; Reitze et al. 2019a) and the Einstein Telescope (ET; Punturo et al. 2010). These future observatories will offer unprecedented sensitivity, capable of detecting compact binary coalescences from the earliest epochs of cosmic history. They will routinely observe events with extraordinarily high signal-to-noise ratios (SNRs) reaching several thousand. This leap in sensitivity – along with a redshift reach an order of magnitude beyond current detectors – will open new windows into precision cosmology, tests of gravity, and astrophysical models of binary formation and evolution (Abac et al. 2025).

However, with this increased detection capability come new challenges. One of the most pressing is the overlap of multiple GW signals in the time domain due to the high event rate (Baibhav et al. 2019). When signals overlap in time or frequency, standard data analysis pipelines may no longer be reliable, potentially introducing significant biases in the inferred source parameters.

In our exploratory study, we were among the first (see also the independent analysis by Samajdar et al. 2021) to quantify these biases by testing existing parameter inference pipelines in the presence of overlapping GW signals. We simulate various configurations of two overlapping signals from non-spinning binaries, systematically varying their relative SNRs, coalescence times, and merger phases. We show that – by setting a prior on the coalescence time using the information from detection pipelines, which are typically accurate to within ~ 10 ms (Regimbau et al. 2012; Meacher et al. 2016) – it is possible to correctly infer the properties of multiple overlapping signals even with the current data-analysis infrastructure, provided that the coalescence times of the signals in the detector frame are more than $\sim 1-2$ seconds apart. However, if the coalescence times differ by less than ~ 0.5 seconds, significant biases arise, highlighting the need for new analysis strategies and algorithms (e.g., Baka et al. 2025).

Chapter 7 shifts focus to the physics of protoplanetary discs. As in the case of accretion discs around SMBHs, gas turbulence plays a central role in driving accretion and the secular evolution of protoplanetary discs. However, in this context, its influence extends well beyond accretion alone – it affects a wide range of processes that are crucial for planet formation. Quantifying the level of gas turbulence in discs is therefore one of the key open questions in the field (Rosotti 2023).

A promising approach to constraining turbulence is to measure the vertical scale height of the dust layer in discs, which is expected to trace the gas structure through gas–dust coupling. This has become feasible thanks to the unprecedented resolution of ALMA observations, which have revealed detailed substructures – such as rings and gaps – in the 2D emission profiles of protoplanetary discs (Bae et al. 2022). As shown by Pinte et al. (2016), it is possible to exploit these features to uncover the 3D morphology of discs. The idea is simple: due to projection effects, a gap in a disc's emission profile will be partly filled by the emission coming from the neighbouring regions. This effect is stronger along the minor axis of the disc, whereas the major axis is only marginally affected. Hence, one can compare the gap contrast along the major and minor axes to infer the degree of this "filling", which in turn depends on the disc's vertical thickness.

In our study, we applied this technique to high-resolution ALMA data from the DSHARP survey (Andrews et al. 2018), constructing radiative transfer models to reproduce the observed gap contrast for varying dust scale heights. We find that, in discs where constraints are possible, the preferred models favor small scale heights, indicating low levels of gas turbulence. For

the remaining nine systems in our sample, our method yields no meaningful constraints, likely due to either low disc inclination or insufficiently deep gaps. Based on our analysis, we propose an empirical criterion to assess whether a given disc is suitable for this technique, offering a valuable tool for guiding future observational efforts.

2 Revisiting the extreme clustering of $z \approx 4$ quasars with large volume cosmological simulations

Abstract

Observations from wide-field quasar surveys indicate that the quasar auto-correlation length increases dramatically from $z \approx 2.5$ to $z \approx 4$. This large clustering amplitude at $z \approx 4$ has proven hard to interpret theoretically, as it implies that quasars are hosted by the most massive dark matter halos residing in the most extreme environments at that redshift. In this work, we present a model that simultaneously reproduces both the observed quasar auto-correlation and quasar luminosity functions. The spatial distribution of halos and their relative abundance are obtained via a novel method that computes the halo mass and halo cross-correlation functions by combining multiple large-volume dark-matter-only cosmological simulations with different box sizes and resolutions. Armed with these halo properties, our model exploits the conditional luminosity function framework to describe the stochastic relationship between quasar luminosity, L, and halo mass, M. Assuming a simple power-law relation $L \propto M^{\gamma}$ with log-normal scatter, σ, we are able to reproduce observations at $z \sim 4$ and find that: (a) the quasar luminosity-halo mass relation is highly non-linear ($\gamma \gtrsim 2$), with very little scatter ($\sigma \lesssim 0.3$ dex); (b) luminous quasars ($\log_{10} L/\mathrm{erg\,s^{-1}} \gtrsim 46.5 - 47$) are hosted by halos with mass $\log_{10} M/\mathrm{M}_{\odot} \gtrsim 13 - 13.5$; and (c) the implied duty cycle for quasar activity approaches unity ($\varepsilon_{\mathrm{DC}} \approx 10 - 60\%$). We also consider observations at $z \approx 2.5$ and find that the quasar luminosity-halo mass relation evolves significantly with cosmic time, implying a rapid change in quasar host halo masses and duty cycles, which in turn suggests concurrent evolution in black hole scaling relations and/or accretion efficiency.

Published in: **EP**, Joseph F Hennawi, Joop Schaye, Matthieu Schaller, *Revisiting the extreme clustering of $z \approx 4$ quasars with large volume cosmological simulations*, Monthly Notices of the Royal Astronomical Society, Volume 528, Issue 3, March 2024, Pages 4466–4489, doi.org/10.1093/mnras/stae329
Reprinted here in its entirety.

2.1 Introduction

Quasars are extreme manifestations of the supermassive black holes (SMBHs) that are thought to reside at the center of almost every massive galaxy (e.g., Salpeter 1964; Zel'dovich & Novikov 1967; Lynden-Bell 1969; Magorrian et al. 1998; Ferrarese & Merritt 2000; Kormendy & Ho 2013). Investigating the characteristics of these luminous objects has been an active area of research for more than half a century (Schmidt 1963). In the last few years, it has become possible to trace their evolution up to redshift $z \approx 7$ (Yang et al. 2020; Bañados et al. 2018; Wang et al. 2021; see also Fan et al. 2023 for a review). Understanding the properties of quasars such as their abundance, luminosity, and spatial distribution, as well as their evolution with redshift, is a key step in order to study the interplay between supermassive black holes, their host galaxies, and the intergalactic medium (IGM) over cosmic time.

In particular, measuring the clustering of quasars is crucial for gaining information on the large-scale environment in which these objects reside. Like their host halos, quasars are biased tracers of the underlying distribution of dark matter (e.g., Kaiser 1984; Bardeen et al. 1986). For this reason, obtaining an estimate for the linear bias factor of quasars (e.g., by measuring the quasar auto-correlation function) makes it possible to infer the characteristic masses of the halos hosting active quasars. In turn, these masses can shed light on the large-scale environment that quasars inhabit, and – by comparing the number density of quasars with that of the hosting halos – on the fraction of time SMBHs are shining as active quasars (known as the quasar duty cycle; see e.g. Martini & Weinberg 2001; Haiman & Hui 2001; Martini 2004).

Thanks to large-sky surveys such as the Sloan Digital Sky Survey (SDSS, York et al. 2000) and the 2dF QSO redshift survey (2QZ, Croom et al. 2004), measurements of large-scale quasar clustering up to $z \approx 4$ have been available for more than a decade. However, a satisfactory theoretical interpretation of these data at all redshifts is still lacking. This is mainly due to the surprising evidence that the bias factor of quasars is a steep function of redshift (Porciani et al. 2004; Croom et al. 2005; Porciani & Norberg 2006; Shen et al. 2007; Ross et al. 2009; Eftekharzadeh et al. 2015; McGreer et al. 2016; Yue et al. 2021; Arita et al. 2023). While in the local universe quasars trace halos in a way that is similar to optically selected galaxies, with a bias factor close to unity (Croom et al. 2005; Ross et al. 2009), at $z \approx 4$ they are the most highly clustered objects known at that epoch, with a bias factor as high as $b \approx 15$ (or, equivalently, a quasar auto-correlation length of $r_0 \approx 24\,\mathrm{cMpc}\,h^{-1}$; Shen et al. 2007, hereafter S07). Such a large correlation length implies that quasars are rare objects, arising only in the most massive halos and shining for a large fraction of the Hubble time.

Several theoretical studies have tried to reproduce the results of S07 at $z \approx 4$. White et al. (2008) developed a simple model for quasar demographics that builds on a linear relation between quasar luminosity and host halos mass. They showed that to match the bias measured in S07, the scatter in this relation must be very small ($\lesssim 0.3$ dex). This conclusion poses two fundamental problems. Firstly, such a low scatter in the quasar luminosity-halo mass relation would be very surprising. In fact, the conventional wisdom on the coevolution of quasars and host galaxies/halos implies that there are multiple sources of scatter contributing to determining the luminosity of a quasar at fixed halo mass (the scatter in the relations between black hole mass and quasar luminosity, black hole mass and bulge mass and between bulge mass and halo mass). A second concern is that low scatter in the luminosity of quasars seems to be in contrast with measurements of the relative abundance of quasars at different luminosities (the so-called quasar luminosity function, QLF). It has long been established that the bright end of the QLF is well-fitted by a power-law (e.g., Boyle et al. 2000; Richards et al. 2006), which stands in contrast with the exponentially-declining halo mass and galaxy luminosity functions (Press & Schechter 1974; Schechter 1976). The easiest way to connect these different shapes is via significant scatter in the luminosity of quasars at fixed halo/galaxy mass. Indeed, a number of demographic models have been developed to interpret the abundance of bright quasars and link them to their host halos (e.g., Croton 2009; Conroy & White 2013; Fanidakis et al. 2013; Veale et al. 2014; Ren et al. 2020; Ren & Trenti 2021; Zhang et al. 2023b). All of these studies (sometimes only implicitly) explain the relatively large number of very luminous quasars by demanding a broad range of possible quasar luminosities at a given host mass so that the more abundant population of lower-mass halos can also host a significant (or even dominant; e.g., Zhang et al. 2023a) fraction of the very bright quasars. As pointed out by some of these same studies, however, the masses of the quasar hosts implied by this picture are in plain contrast with the high masses necessary to account for the S07 bias measurement.

In summary, the very strong clustering measured by S07 implies a very small scatter in the luminosity of quasars at a given halo mass, and this is in tension with the large scatter required by physical models of the quasar luminosity function. A first attempt at solving the tension was made by Shankar et al. (2010b), using a model that connects quasar luminosities and black hole masses while accounting for the growth of black holes during cosmic time in a self-consistent way. The authors of this study tried to match simultaneously the value of the bias inferred by S07 and several measurements of the QLF at $z = 3 - 6$ (Shankar & Mathur 2007; Shankar 2009). Assuming a non-linear relation between halo mass and quasar luminosity, they find that a low value of the scatter in this relation can reproduce the measurements of the bright end of the QLF. Even when assuming that all massive halos contribute to the clustering of quasars (i.e., a quasar duty cycle for massive

systems equal to unity), however, their prediction for the $z = 4$ quasar clustering is ≈ 2 standard deviations below the value measured by S07. Wyithe & Loeb (2009) also find that the S07 bias measurement cannot be reproduced when assuming that the bias of dark matter halos is solely a function of their mass, and suggest that stronger clustering could be obtained if quasar activity was sparked by recent mergers (the so-called "assembly/merger bias", see e.g., Furlanetto & Kamionkowski 2006; Wetzel et al. 2009; see also Wechsler & Tinker 2018). However, Bonoli et al. (2010) (see also Cen & Safarzadeh 2015) used the Millennium Simulation (Springel et al. 2005) to study whether recently merged massive halos were clustered more strongly than other halos of the same mass, but found no evidence for that.

Numerous other studies have compared their predictions for the quasar clustering to the S07 measurements, using a variety of different approaches such as empirical models of quasar-galaxy coevolution (Kauffmann & Haehnelt 2002; Hopkins et al. 2007b; Croton 2009; Shankar et al. 2010a; Conroy & White 2013; Aversa et al. 2015; Shankar et al. 2020), semi-analytic models of galaxy formation (Bonoli et al. 2009; Fanidakis et al. 2013; Oogi et al. 2016) and cosmological hydrodynamical simulations (DeGraf et al. 2012; DeGraf & Sijacki 2017). While these studies are generally successful in reproducing the quasar auto-correlation function (or, equivalently, the quasar linear bias) at lower redshift ($z \lesssim 3$), none of these studies have been shown to be compatible with the strong clustering observed by S07.

In conclusion, despite the efforts that have been devoted to interpreting the auto-correlation function of quasars at high redshift, a number of questions remain open: (a) is the S07 measurement compatible with the standard cosmological model in which clustering is dictated by halo mass or is something akin to assembly bias playing an important role? (b) What is the scatter in the quasar luminosity-halo mass relation? Can small (large) scatter be reconciled with the observed quasar luminosity function (auto-correlation function)? (c) What are the physical properties that can be inferred from jointly modeling the QLF and quasar clustering? Can the characteristic mass of host halos and the quasar duty cycle be determined precisely?

One of the reasons why we have not been able to give definitive answers to these questions in more than a decade, is that modeling the clustering of high redshift quasars is difficult. The works of White et al. (2008), Shankar et al. (2010b), and Wyithe & Loeb (2009) clearly show that the results of their theoretical models are strongly dependent on the assumed functional form for the linear bias-halo mass relation. This is because the different analytical predictions for this relation based on linear theory (e.g. Mo & White 1996; Jing 1998; Sheth et al. 2001) diverge significantly at masses that correspond to peaks in the density perturbations that are already very non-linear (Barkana & Loeb 2001). For the case considered here, a bias of $b \approx 15$,

i.e., the value measured by S07 for $z \approx 4$ quasars, corresponds to a value of the peak height $\nu = \delta_c/\sigma(M, z)$ – with $\delta_c \approx 1.69$ and $\sigma^2(M, z)$ being the variance of the smoothed linear density field – equal to $\nu \approx 4 - 6$, depending on the specific linear bias-halo mass relation and cosmology considered. Such values are rather extreme, implying that the systems contributing to the clustering of $z = 4$ quasars live in very rare and massive environments that depart very early from the behavior expected for a linear density field.

Improving the accuracy of the linear bias-halo mass relation via empirical fits to cosmological N-body simulations (e.g., Shankar et al. 2010b; Tinker et al. 2010; Comparat et al. 2017) does not provide a complete solution to the problem. In fact, the key point here is that the use of the large-scale linear bias as a proxy for the clustering of quasars assumes that the measured data are on quasi-linear scales, where the distribution of quasars is related to the underlying matter distribution by a scale-independent factor. This assumption breaks down for the small scales (as low as $\approx 5\,\mathrm{cMpc}$) and for the highly non-linear environments probed by the S07 data. For the same reason, an approach based on the (analytical) halo model framework (Cooray & Sheth 2002) would also be problematic, as the non-linear bias plays a relevant role in the transition region between the one-halo and the two-halo contributions (e.g., Mead & Verde 2021; Nishimichi et al. 2021).

In this paper, we aim to directly reproduce the observed $z = 4$ quasar auto-correlation function (S07) in its entirety by making use of large-volume cosmological simulations. This is a challenging numerical problem: in order to model the auto-correlation function properly, we need to obtain a large statistical sample of halos with masses up to $M \approx 10^{13} - 10^{14}\,\mathrm{M}_\odot$ (which correspond, at $z = 4$, to the peak heights mentioned above, i.e., $\nu \approx 4 - 6$). Given the fact that the mass function declines exponentially at large masses, these halos are extremely rare ($1 - 10\,\mathrm{cGpc}^{-3}$), and therefore a very large simulated volume of more than $\approx 100\,\mathrm{cGpc}^3$ is needed to obtain a sample of at least $\approx 10^2 - 10^3$ massive halos, that can be used to properly model the quasar auto-correlation function even at the highest masses. This is in agreement with the fact that the comoving volume probed by the SDSS observations used by S07 is around $\approx 50\,\mathrm{cGpc}^3$. A volume larger than the observational one is necessary to build a model for the quasar auto-correlation function that has higher signal-to-noise ratio than the data. At the same time, however, we also want to resolve halos down to $M \approx 10^{11} - 10^{12}\,\mathrm{M}_\odot$ in order to explore the different possible distributions of quasars in halos that can give rise to the observed clustering. To probe these very different halo masses, we make use of a new semi-analytical framework (Sec. 2.2.2.2 and Appendix 2.B) that allows us to employ multiple simulated boxes to widen the range of masses that can be properly modeled by our simulations.

We employ the dark-matter-only versions of the FLAMINGO suite of cosmological simulations (Schaye et al. 2023; Kugel et al. 2023) and focus on two specific box sizes: $L = 2.8\,\mathrm{cGpc}$ and $L = 5.6\,\mathrm{cGpc}$. On top of reproduc-

ing the clustering measurements at $z = 4$, we also consider the constraints coming from the relative abundance of quasars at the same redshift. In other words, we aim to match the observed quasar auto-correlation and luminosity functions simultaneously. We make use of the spatial and mass distribution of halos in the simulated volumes to build a simple quasar population model that can be directly compared with observations. In this way, we are able to investigate the predictive power of quasar observables in a ΛCDM framework and obtain physical constraints on the halo mass distribution of quasar hosts and the quasar duty cycle. We also use our model to analyze the clustering and luminosity function data at a lower redshift ($z \approx 2 - 3$), where the tension between theoretical models and data is not as strong (e.g., Croton 2009; Conroy & White 2013; Aversa et al. 2015). This serves as a benchmark of the validity of our model and allows us to discuss the evolution of the physical properties of quasars with redshift.

The paper is structured as follows. In Sec. 2.2 we discuss the basic assumptions of the model, outline the cosmological simulations employed in our work, and describe how we extract the physical quantities that are necessary to model the quasar correlation function and luminosity function simultaneously. Sec. 2.3 gives a brief overview of the data we compare our model with, and it provides details on the statistical methodology underlying that comparison. Sec. 2.4 presents the main results of our analysis, while Sec. 2.5 contains a discussion of the implications of our findings and their connections to previous work. Conclusions are provided in Sec. 2.6.

2.2 Methods

In this Section, we describe our model for the distribution of quasars in space and luminosity. We start by outlining the basic framework (Sec. 2.2.1); then, we describe the FLAMINGO cosmological simulations and detail how we extract the mass function and the cross-correlation functions of halos (Sec. 2.2.2). Figure 2.1 shows a summary of the various quantities involved in our analysis, together with a reference to the equations where they are defined.

2.2.1 The conditional luminosity function

We adopt an empirical model that is agnostic to the physics underlying the quasar emission/black hole accretion mechanisms. The only assumptions we make are: (a) every halo above some mass $M_{\rm min}$ hosts a SMBH at its center, emitting at some luminosity L; (b) the luminosity of a SMBH depends only on the mass of the host halo, M. Therefore, we can employ a conditional luminosity function approach (CLF; see e.g., Yang et al. 2003; Ren et al.

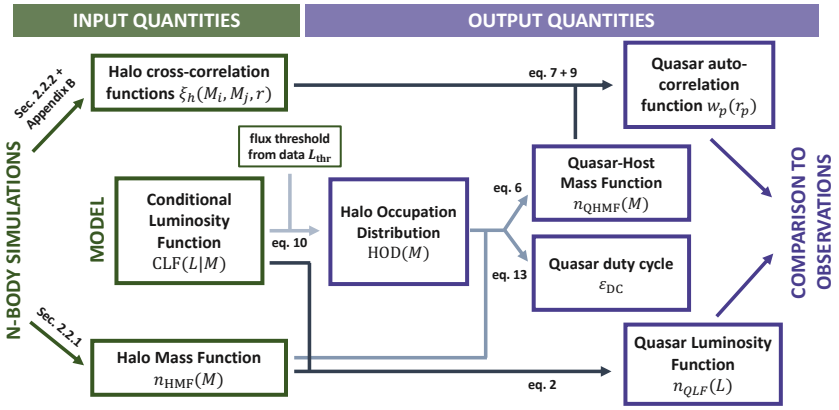

Figure 2.1: Summary of the various quantities involved in the analysis. We choose a model for the Conditional Luminosity Function (CLF) that depends on a set of free parameters. We then combine this with the halo mass function and the halo cross-correlation functions taken from the FLAMINGO cosmological simulations to obtain the two main observables of interest, the quasar luminosity function and auto-correlation function, together with other key properties such as the quasar-host mass function, the halo occupation distribution (HOD), and the quasar duty cycle.

2020) and write the 2-d distribution in the black hole luminosity-host halo mass plane, $n(L, M)$, as:

$$n(L, M) = \mathrm{CLF}(L|M)\, n_{\rm HMF}(M), \tag{2.1}$$

where $n_{\rm HMF}(M)$ is the halo mass function.

Note that the luminosity of a SMBH, L, can be interpreted as either a bolometric luminosity or a luminosity in a specific band of the spectrum. The framework that we are introducing here is agnostic to this choice and can be formulated to describe the emission coming from any region of the spectrum. However, for clarity and consistency with previous work on the subject (e.g., White et al. 2008; Shankar et al. 2010b; Conroy & White 2013; Zhang et al. 2023b), in this paper we choose to work with bolometric luminosities. Henceforth, L will always refer to the bolometric luminosity, i.e., $L \equiv L_{\rm bol}$[1].

Within this framework, the Quasar Luminosity Function – $n_{\rm QLF}(L)$ – is simply the marginalization of $n(L, M)$ over halo mass, M:

$$n_{\rm QLF}(L) = \int_{M_{\rm min}}^{M_{\rm max}} \mathrm{CLF}(L|M)\, n_{\rm HMF}(M)\, \mathrm{d}M. \tag{2.2}$$

[1]However, note that the data considered in this paper always refer to type I, UV-bright quasars (e.g., Padovani et al. 2017). Hence, the model presented in this work describes only this specific population of active SMBHs.

Therefore, assuming that the halo mass function is known, the QLF can be easily determined once a conditional luminosity function has been adopted. The two limits of integrations, $M_{\rm min}$ and $M_{\rm max}$, represent the minimum/maximum mass of a halo that can host a SMBH. In principle, we could have SMBHs in any halos, and set this integration range to be as wide as possible. However, given that the simulations employed in our analysis span a wide but finite range of masses (Sec. 2.2.2), we adopt the following limits: $\log_{10} M_{\rm min}/{\rm M}_\odot = 11.5$, and $\log_{10} M_{\rm max}/{\rm M}_\odot = 14$ ($\log_{10} M_{\rm max}/{\rm M}_\odot = 14.5$) at redshift $z = 4$ ($z = 2.5$). These limits enclose a range in masses that is sufficiently broad for our redshifts of interest (Sec. 2.4), so that expanding the range would have a negligible impact on our final results.

We use a model for the CLF in which the distribution in luminosity is log-normal at fixed mass (see also Ren et al. 2020; Ren & Trenti 2021):

$$\mathrm{CLF}(L|M)\,\mathrm{d}L = \frac{f_{\rm on}}{\sqrt{2\pi}\sigma}\,\exp\left(\frac{(\log_{10} L - \log_{10} L_{\rm c}(M))^2}{2\sigma^2}\right)\mathrm{d}\log_{10} L. \quad (2.3)$$

We then assume a power-law dependence of the characteristic luminosity, $L_{\rm c}$, on mass:

$$L_{\rm c}(M) = L_{\rm ref}\left(\frac{M}{M_{\rm ref}}\right)^{\gamma}, \quad (2.4)$$

or, in terms of logarithmic quantities:

$$\log_{10} L_{\rm c}(M) = \log_{10} L_{\rm ref} + \gamma\,(\log_{10} M - \log_{10} M_{\rm ref})\,, \quad (2.5)$$

where $M_{\rm ref}$ is simply a reference mass that is associated with the reference luminosity $L_{\rm ref}$. We fix $\log_{10} M_{\rm ref}/{\rm M}_\odot = 12.5$. The free parameters of the model are: σ, $L_{\rm ref}$, γ, and $f_{\rm on}$. In the following, we assume that these parameters do not depend on the other variables such as halo mass or quasar luminosity, and let them assume different values for the different redshifts we consider in Sec. 2.4.

The factor $f_{\rm on}$ accounts for the fact that not all black holes may be active as quasars at any given time. Therefore, we are implicitly assuming that the CLF is bimodal: the first mode accounts for all luminous quasars and is log-normally distributed, whereas the second mode (not accounted for in eq. 2.3) describes the behavior of the black holes that are too dim to be probed by any observations and is therefore completely irrelevant to our analysis. This bimodality in the CLF has a well-defined physical meaning: black holes are either active as luminous quasars or they are dormant, with a luminosity that is orders of magnitudes lower than any observational limits. However, it is not clear whether the luminosity distribution of black holes is indeed bimodal, or rather shows a continuum between active sources and inactive/faint ones. Observations of very faint quasars ($\log_{10} L/{\rm erg\,s^{-1}} \approx 42 - 45$) can shed light on this question[2]. We will return to this point in Sec. 2.5.3.

[2]Such observations become very difficult in the distant universe, as faint quasars are often outshined by their host galaxies.

2.2.1.1 The quasar auto-correlation function

In our framework, the correlation function of quasars is identical to the correlation function of the halos that host them, as quasars are temporally subsampling the underlying halo distribution. However, we have to consider that only quasars above some luminosity threshold $L_{\rm thr}$ are accounted for when measuring the correlation function in a survey. Therefore, we are effectively considering a "biased" halo mass distribution traced by the quasars above this luminosity threshold: we will refer to it as the "Quasar-Host Mass Function" (QHMF). This quantity can be expressed in terms of the halo mass function and another marginalization integral of the CLF:

$$n_{\rm QHMF}(M|L > L_{\rm thr}) = n_{\rm HMF}(M) \int_{L_{\rm thr}}^{\infty} {\rm CLF}(L|M)\,{\rm d}L. \tag{2.6}$$

The clustering of quasars can then be determined by computing the correlation function of a sample of halos that are distributed according to $n_{\rm QHMF}(M|L > L_{\rm thr})$. Here, we use an approach that allows us to quickly determine the quasar auto-correlation function for different $n_{\rm QHMF}(M)$ distributions: we create different mass bins, and – selecting halos in these bins – extract the cross-correlation functions for halos with different masses from a cosmological simulation (see Sec. 2.2.2 for more details). Let us call these cross-correlation terms $\xi_h(M_j, M_k; r)$, with $M_{j,k}$ being the centers of the mass bins. We can then compute the quasar auto-correlation function, $\xi(r)$, by simply weighting the cross-correlations terms, $\xi_h(M_j, M_k; r)$, according to the quasar-host mass function, $n_{\rm QHMF}$:

$$\xi(r) = \sum_{j,k} p_j p_k \xi_h(M_j, M_k; r), \tag{2.7}$$

where the weights $p_{j,k}$ are defined as:

$$p_i = \frac{n_{\rm QHMF}(M_i|L > L_{\rm thr})\,\Delta M}{\int_0^{M_{\rm max}} n_{\rm QHMF}(M|L > L_{\rm thr})\,{\rm d}M}, \tag{2.8}$$

with ΔM being the width of the mass bins. We present how to derive these equations in Appendix 2.A.

Once $\xi(r)$ is known, other related quantities such as the projected auto-correlation function, $w_{\rm p}(r_{\rm p})$, can be easily obtained by integrating along the parallel direction π. The projected auto-correlation function is relevant since it can be directly compared to observational data (see Sec. 2.3.1). Setting a maximum value for the parallel distance $\pi_{\rm max}$, which is chosen in accordance with the one used for observational data, e.g. $\pi_{\rm max} = 100\,{\rm cMpc}\,h^{-1}$ for the S07 measurements, the projected auto-correlation function reads:

$$w_{\rm p}(r_{\rm p}) = \int_{-\pi_{\rm max}}^{\pi_{\rm max}} \xi(r_{\rm p}, \pi)\,{\rm d}\pi = 2\int_{r_{\rm p}}^{\sqrt{r_{\rm p}^2+\pi_{\rm max}^2}} \frac{r\xi(r)}{\sqrt{r^2 - r_{\rm p}^2}}\,{\rm d}r. \tag{2.9}$$

2.2.1.2 Halo occupation distribution and duty cycle

From the CLF, we can extract other quantities that will be relevant to our analysis. In particular, the integral of the CLF above some threshold luminosity $L_{\rm thr}$ represents the aggregate probability for a halo of mass M to host a quasar with a luminosity above the threshold value. Therefore, it is equivalent to a Halo Occupation Distribution (HOD; see e.g., Berlind & Weinberg 2002):

$$\mathrm{HOD}(M) = \frac{n_{\rm QHMF}(M|L > L_{\rm thr})}{n_{\rm HMF}(M)} = \int_{L_{\rm thr}}^{\infty} \mathrm{CLF}(L|M)\,\mathrm{d}L. \tag{2.10}$$

The HOD is also closely related to the idea of a *quasar duty cycle*. In fact, the duty cycle is defined as the fraction of active quasars (i.e., with a luminosity above the threshold) divided by the fraction of halos that are able to host these quasars. In the standard picture (e.g., Martini & Weinberg 2001; Haiman & Hui 2001) this fraction is well defined, as it is implicitly assumed that there is a minimum halo mass $\tilde{M}_{\rm min}$ above which all halos can host quasars, and only a fraction of them is active at the present moment. In other words, the QHMF is:

$$n_{\rm QHMF}(M) = \varepsilon_{\rm DC} n_{\rm HMF}(M) \Theta(\log_{10} M - \log_{10} \tilde{M}_{\rm min}), \tag{2.11}$$

with $\varepsilon_{\rm DC}$ being the duty cycle and Θ the Heaviside step function. However, this definition of the duty cycle is not well-posed in our approach. As described above, we do not assume a specific functional form for the QHMF, but rather we infer this quantity from the CLF (eq. 2.6). As a consequence, we do not define a minimum mass $\tilde{M}_{\rm min}$ for halos to host bright quasars, but rather consider all halos and compute a probability for them to host bright quasars given their mass, M. This implies that, in principle, even halos with a very low mass could have a small but non-zero probability to host bright quasars. As a result, the above definition of the quasar duty cycle would return artificially small values. Therefore, we opt here for an alternative definition (see also Ren & Trenti 2021): the duty cycle, $\varepsilon_{\rm DC}$, is the weighted average of the HOD above a threshold mass that is given by the median of the QHMF, $n_{\rm QHMF}(M|L > L_{\rm thr})$. In other words, if we define the median of the QHMF as the mass $M_{\rm med}$ satisfying the relation:

$$\int_{M_{\rm med}}^{M_{\rm max}} n_{\rm QHMF}(M) = 0.5 \int_{M_{\rm min}}^{M_{\rm max}} n_{\rm QHMF}(M), \tag{2.12}$$

then $\varepsilon_{\rm DC}$ can be expressed as:

$$\varepsilon_{\rm DC} = \frac{\int_{M_{\rm med}}^{M_{\rm max}} n_{\rm HMF}(M)\,{\rm HOD}(M)\,{\rm d}M}{\int_{M_{\rm med}}^{M_{\rm max}} n_{\rm HMF}(M)\,{\rm d}M} = \\ = \frac{\int_{M_{\rm med}}^{M_{\rm max}} n_{\rm QHMF}(M|L > L_{\rm thr})\,{\rm d}M}{\int_{M_{\rm med}}^{M_{\rm max}} n_{\rm HMF}(M)\,{\rm d}M}. \tag{2.13}$$

2.2.2 Dark matter only simulation setup

In the last section, we have shown how we can make use of the CLF formalism to compute the quasar luminosity and auto-correlation functions – together with other relevant quantities such as the QHMF and the quasar duty cycle – using two fundamental ingredients: the mass function of halos and the cross-correlation functions of halos with different masses (see Figure 2.1 for a summary of this workflow). In this section, we provide details on how we obtain these two ingredients using the Dark-Matter-Only (DMO) version of the FLAMINGO suite of cosmological simulations.

FLAMINGO (Schaye et al. 2023; Kugel et al. 2023) is a suite of state-of-the-art, large-volume cosmological simulations run with the N-body gravity and smooth particle hydrodynamics (SPH) solver SWIFT (Schaller et al. 2024). Gravity is solved using the Fast Multiple Method (Greengard & Rokhlin 1987). The cosmology adopted in FLAMINGO is the "3x2pt + all" cosmology from Abbott et al. (2022) ($\Omega_{\rm m} = 0.306$, $\Omega_{\rm b} = 0.0486$, $\sigma_8 = 0.807$, $\rm H_0 = 68.1\,km\,s^{-1}\,Mpc^{-1}$, $n_{\rm s} = 0.967$), with a summed neutrino mass of 0.06 eV. Massive neutrinos are included in the simulation via the δf method of Elbers et al. (2021). Initial conditions (ICs) are set using multi-fluid third-order Lagrangian perturbation theory (3LPT). Partially fixed ICs are used to limit the impact of cosmic variance (Angulo & Pontzen 2016) by setting the amplitudes of modes with $(kL)^2 < 1025$ to the mean variance (k is the wavenumber and L the box size).

In this work, we focus on two specific DMO simulations with box sizes $L = 2800\,{\rm cMpc}$ and $L = 5600\,{\rm cMpc}$, respectively. Both simulations have 5040^3 cold dark matter (CDM) particles and 2800^3 neutrino particles. The CDM particle masses are $M_{\rm dm} = 6.72 \times 10^9\,{\rm M_\odot}$ and $M_{\rm dm} = 5.38 \times 10^{10}\,{\rm M_\odot}$ for the $L = 2800\,{\rm cMpc}$ and $L = 5600\,{\rm cMpc}$ boxes, respectively. We focus on the DMO version of the simulations because no hydrodynamic version is available for the largest box, and because we are only interested in the spatial distribution of halos, that, in the ΛCDM model, is primarily dictated by gravitational interactions of dark-matter particles only.

We identify halos in the simulated snapshots using the 6-d friends-of-friends code VELOCIRAPTOR (Elahi et al. 2019). Once halos have been

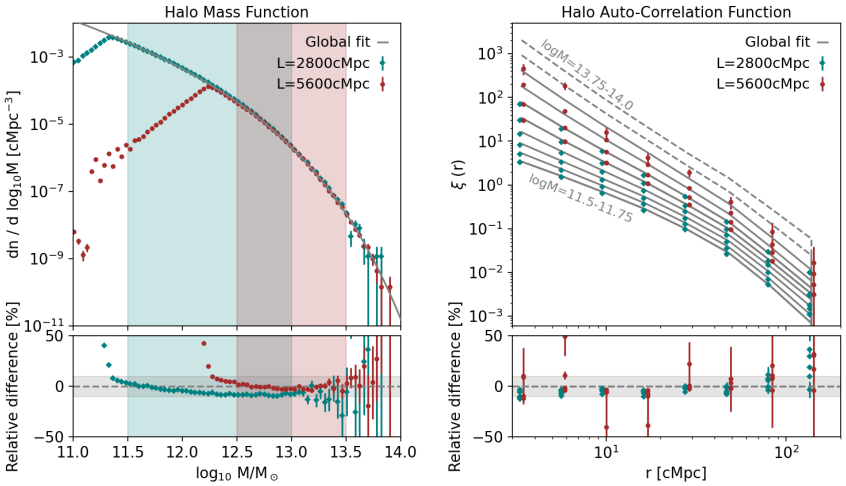

Figure 2.2: *Left:* Halo mass function at $z = 4$ from the simulations considered: $L = 2800\,\mathrm{cMpc}$ (teal diamonds) and $L = 5600\,\mathrm{cMpc}$ (red circles). The gray solid line represents the analytical fit to the simulations (see Sec. 2.2.2.1 for more details). The shaded regions highlight which masses in each simulation are considered for the fit. The bottom panel shows the relative difference between the fit and the two simulations (with the horizontal shaded grey band highlighting the 10% limit). *Right:* Auto-correlation function of halos in different mass bins at $z = 4$. We create 8 mass bins ranging from $\log_{10} M/\mathrm{M}_\odot = 11.5$ to $\log_{10} M/\mathrm{M}_\odot = 13.5$ and 0.25 dex wide. Lower mass bins correspond to lower values of the correlation functions, and vice-versa. Teal diamonds refer to the $L = 2800\,\mathrm{cMpc}$ simulation, while red circles refer to the $L = 5600\,\mathrm{cMpc}$ one. Points are staggered in the x-direction for visualization purposes. The gray solid lines represent the fits to the auto-correlation functions (from the lowest mass bin on the bottom to the highest mass bin on top), as described in Appendix 2.B. Relative differences between the fits and the simulated correlation functions are shown in the bottom panel. These differences are generally $\lesssim 10\%$, with the exception of the highest mass bin considered (i.e., $\log_{10} M/\mathrm{M}_\odot = 13.25 - 13.5$), for which the measurements are noisy due to the small number of halos in that mass range. The gray dashed lines in the top panel show extrapolations of the auto-correlation functions based on our fit for even higher mass bins ($\log_{10} M/\mathrm{M}_\odot = 13.5 - 13.75$ and $\log_{10} M/\mathrm{M}_\odot = 13.75 - 14.0$) where measurements from the simulations are not available. More details can be found in Sec. 2.2.2.2 and Appendix 2.B.

identified, their masses are computed using a spherical-overdensity definition based on their density profile. We perform this task using the code SOAP[3]. We define the radius of a halo as the distance from the most bound particle within which the density reaches a value of 200 times the critical density of the universe ($200\rho_{\rm c}$). We only include central halos in the analysis and exclude the contribution of sub-halos. As discussed in Sec. 2.5.3, we do not expect this to influence our results significantly.

Once we have obtained a catalogue with the positions and masses of halos in the simulation at a given redshift, we can easily compute key statistical properties such as the halo mass function and the (cross-)correlation functions of halos with different masses. However, this approach is not directly suitable for our purposes. In fact, an important limitation of cosmological simulations is that they give reliable results only in a finite range of masses. The lower limit of this mass range is imposed by resolution: halos with fewer than $50-100$ dark-matter particles are not well resolved, and thus cannot be trusted. The upper limit, on the other hand, is set by the box size of the simulation: if the number of halos with mass greater than some threshold M is small, these halos are too rare to get a reliable estimate of their statistical properties (e.g., their clustering).

For the problem we are facing here, we need to be able to reproduce the relative abundance of halos and their spatial distribution for a vast range of masses. For this reason, employing a single halo catalogue obtained using a simulation with a fixed box size is not the optimal strategy. Instead, we use here an approach consisting of two key steps: we first compute the quantities of interest (i.e., the halo mass function and the halo cross-correlation functions) from multiple simulations with different box sizes (and mass resolutions), and then we combine these different simulations by making use of analytical fitting functions. In this way, we can predict the abundance and spatial distribution of halos for all the masses that are well captured by the different simulations considered.

Table 2.1 summarizes the properties of the simulations we employ. In brief, we use the two different box sizes $L = 2800\,{\rm cMpc}$ and $L = 5600\,{\rm cMpc}$ to study the properties of low-mass and high-mass halos, respectively. For the $2800\,{\rm cMpc}$ box, we select halos in the range of masses $\log_{10} M/{\rm M}_\odot = 11.5-13.0$; for $L = 5600\,{\rm cMpc}$, we focus on halos in the range $\log_{10} M/{\rm M}_\odot = 12.5-13.5$. The lower limits are set to select only halos with at least ≈ 50 particles, whereas the upper limits are set to ensure overlap between the two mass ranges and to guarantee that all mass bins (up to at least $z = 4$) are populated with at least 5000 halos. In the following we describe in detail how we combine these simulations to obtain an analytical description of the halo mass function and of the cross-correlation function of halos with different masses.

[3] https://github.com/SWIFTSIM/SOAP

Table 2.1: Summary of the different FLAMINGO cosmological simulations employed in the analysis. The "fitting mass range" refers to the mass range selected for the fits of the halo mass function and the cross-correlation functions (Sec. 2.2.2.1 and 2.2.2.2, respectively). The redshifts considered in the analysis are $z = 4.0$ (high-redshift data; see Figure 2.2), and $z = 2.5$ (low-redshift data; see Figure 2.9).

Box size [cMpc]	Number of CDM particles	CDM particle mass [$\log_{10} \mathrm{M}_\odot$]	Fitting mass range [$\log_{10} \mathrm{M}_\odot$]	Snapshots considered (z)
2800	5040^3	9.83	$11.5 - 13.0$	2.5, 4.0
5600	5040^3	10.77	$12.5 - 13.5$	2.5, 4.0

2.2.2.1 Fitting the halo mass function

Following Tinker et al. (2008) (see also Jenkins et al. 2001; White 2001; Warren et al. 2006), we write the halo mass function in terms of the peak height of the density perturbations, $\nu = \delta_c/\sigma(M,z)$, where $\delta_c \approx 1.69$ is the critical linear density for collapse and $\sigma(M,z)$ is the variance of the linear density field smoothed on a scale $R(M)$ (see Press & Schechter 1974; Sheth & Tormen 1999). According to this formalism, the mass function can be parametrized in terms of a universal function $f(\sigma)$:

$$f(\sigma; A, a, b, c) = A \left(\left(\frac{\sigma}{b} \right)^{-a} + 1 \right) e^{-c/\sigma^2}, \tag{2.14}$$

where $f(\sigma)$ is related to the mass function via the expression

$$\frac{\mathrm{d}n}{\mathrm{d}M}(M,z) = f(\sigma) \frac{\rho_{\mathrm{m},0}}{M} \frac{\mathrm{d}\ln\sigma^{-1}}{\mathrm{d}M}, \tag{2.15}$$

with $\rho_{\mathrm{m},0}$ being the mass density at $z = 0$.

We use the python package COLOSSUS (Diemer 2018) to compute the value of $\sigma(M,z)$ using the same cosmology as the FLAMINGO simulation (Sec. 2.2.2). We then use χ^2-minimization to find the best-fitting parameters (A, a, b, c) for the analytical form of the halo mass function. We fit the number density of halos in different mass bins using halo catalogues from two different simulations, using two different (but partially overlapping) mass ranges (see Table 2.1). We assign Poissonian counting errors to every mass bin considered. We also experiment with changing these errors, and find that we achieve a better fit to the data by doubling the errors for the $L = 2800\,\mathrm{cMpc}$ simulation, and halving the ones associated with the $L = 5600\,\mathrm{cMpc}$ box. Note that this choice is arbitrary: our goal is not to provide a physically-motivated fit to the data, but simply to find a good analytical description of the halo mass function coming from simulations.

Figure 2.2 (left panel) shows the best-fitting mass function for $z = 4$, together with the data obtained from the simulations. Analogous results for $z = 2.5$ are shown in Appendix 2.C. The optimal parameter values for this mass function are: $A = 5.68 \times 10^{-5}$, $a = 1.65$, $b = 257$, $c = 1.16$. As shown in the lower left panel of Fig. 2.2, the fit provides a description of the simulated data with an accuracy of $\approx 5-10\%$ up to $\log_{10} M/\mathrm{M}_\odot \lesssim 13.5$. As we will discuss in Sec. 2.5.3, this level of accuracy for the model is enough to provide a satisfactory description of the observed data.

Finally, we note that the reason why we have performed the fitting of the halo mass functions extracted from our simulations and did not consider the best-fitting parameters provided by Tinker et al. (2008) is because we found that, at $z \geq 4$, differences between the Tinker et al. (2008) model and our simulations were as high as 100% (see also Yung et al. 2023).

2.2.2.2 Obtaining the cross-correlation functions

We want to obtain the cross-correlation functions of halos with masses M_j and M_k, $\xi_h(M_j, M_k; r)$. In order to achieve this, we create a grid in mass and distance by considering 8 uniformly spaced bins in $\log_{10} M$, with a minimum halo mass of $\log_{10} M_{\rm min}/{\rm M}_\odot = 11.5$ and a maximum of $\log_{10} M_{\rm max}/{\rm M}_\odot = 13.5$, and 8 (logarithmically-spaced) bins in the radial direction with a minimum radial distance of $\log_{10} r_{\rm min}/{\rm cMpc} = 0.4$ and a maximum of $\log_{10} r_{\rm max}/{\rm cMpc} = 2.25$ We then use the package CORRFUNC (Sinha & Garrison 2020) to compute the number of halo pairs in the simulated catalogues for every combination of masses and distance, together with the number of pairs obtained assuming that these halos are distributed randomly. The values of the cross-correlation terms are obtained using the Landy & Szalay (1993) estimator:

$$\xi_h(M_j, M_k; r) = \xi_{j,k}(r) = \frac{D_j D_k - D_j R_k - D_k R_j + R_j R_k}{R_j R_k}, \qquad (2.16)$$

where $D_j D_k$ stands for the number of pairs of halos in the mass bin j with halos in the mass bin k, whereas $R_j D_k$, $D_j R_k$, and $R_j R_k$ refer to the number of pairs when comparing to a random distribution of the same halos.

We end up with 36 different cross-correlation functions – i.e., the number of independent elements for a symmetric 64-element matrix – which can be used to determine the quasar auto-correlation function according to eq. 2.7. However, once again, we must account for the fact that different simulations probe different mass ranges. We thus fit a parametric analytical function to these cross-correlation functions in a way that allows us to combine different simulated boxes.

Furthermore, in this case the fitting procedure has another critical purpose. Despite the large volume of the simulations employed, the number of simulated halos at the very high mass end is limited by the finite size of the box. For this reason, the obtained cross-correlation terms for the very high-mass halo pairs will suffer from significant uncertainties due to the limited sample size in the simulation. Even for the largest box we consider (i.e., $L = 5600\,{\rm cMpc}$), at $z = 4$ this effect starts to be significant for $\log_{10} M/{\rm M}_\odot \approx 13.2 - 13.5$. This is an important limitation for our analysis: in the inference routine we will undertake in the next Section, we want to be able to explore the full parameter space and consider models for which this range of masses (or even higher) plays a significant role. For this reason, we fit the cross-correlation terms with two key objectives: reducing the noise associated with the poor statistics at the high mass end of the halo mass function, and providing a means to sensibly extrapolate the behavior of the cross-correlation functions up to $\log_{10} M/{\rm M}_\odot = 14.0$ ($\log_{10} M/{\rm M}_\odot = 14.5$) at $z = 4$ ($z = 2.5$). This extrapolation allows us to recover well-behaved posterior distributions (see Sec. 2.4) that provide a complete description

of the different models described by our parameters. Its validity and the associated caveats are discussed in detail in Sec. 2.5.3

We provide details on the fitting of the cross-correlation terms $\xi_h(M_j, M_k; r)$ in Appendix 2.B. In short, we divide all the cross-correlation terms, $\xi_h(M_j, M_k; r)$, by a reference correlation function, $\xi_{\rm ref}(r)$, and fit the results with a 3-d polynomial to capture the residual dependencies on the two masses and on radius. In the rest of this Section, we show the results of the fits for the auto-correlation functions in different mass bins at $z = 4$ (Figure 2.2, right panel; the same plot for $z = 2.5$ is shown in Appendix 2.C). In other words, we plot the correlation functions for bins of equal mass, $\xi_h(M_j, M_j; r)$, together with the fits that are meant to reproduce these functions, $\xi_{h,\rm fit}(M_j, M_j; r)$ (gray lines)[4]. Lower mass bins correspond to lower values of the auto-correlation functions, and vice-versa.

We assign errors to the $\xi_h(M_j, M_j; r)$ points based on the Poissonian statistics of the pair counts; note that in this way we are underestimating the real uncertainties on the data points because we are not including the effects of cosmic variance and of other sources of systematics. For this reason, when assessing the robustness of our fits, it makes little sense to discuss them in terms of statistical errors. We therefore compare the simulated data and the model fits in terms of relative differences between the two (lower right panel of Figure 2.2). These differences are generally at the level of $\lesssim 10\%$ for all bins but the highest one (i.e., $\log_{10} M/{\rm M}_\odot = 13.25 - 13.5$), which is easily recognizable because it has the largest Poissonian uncertainties. As already mentioned before, at very large masses correlation measurements from simulations become noisy (and thus unreliable) due to the small number of halos in the snapshots. Even in this extreme case, however, the fit provides a satisfactory description of the shape and normalization of the correlation function in the simulations, with a relative difference that is still smaller than the uncertainties on the S07 observed data (which are at the level of $50 - 100\,\%$; see Sec. 2.3.1).

Using dashed grey lines, we also plot in Fig. 2.2 the auto-correlations functions for the two bins $\log_{10} M/{\rm M}_\odot = 13.5 - 13.75$ and $\log_{10} M/{\rm M}_\odot = 13.75 - 14.0$, as obtained by extrapolating our fitting functions to masses higher than the ones probed by the simulations. We see that the trend of the auto-correlation functions with halo mass is well preserved by these extrapolations; further discussion on this can be found in Sec. 2.5.3 and Appendix 2.C.

Finally, we note that relative differences between our fits and the values of correlation functions extracted from simulations tend to be larger at very large scales ($r \gtrsim 100\,{\rm cMpc}$). This is also due to the fact that simulation-based values become less reliable in this regime. There are two reasons for that:

[4]The global fits to all the cross-correlation terms $\xi_h(M_j, M_k; r)$ at both redshifts are shown in Appendix 2.B.

first, the finite size of the box reduces the number of very large-scale pairs that are available. Secondly, at $r \gtrsim 100\,\mathrm{cMpc}$ the behavior of correlation functions becomes non-trivial due to the presence of the baryon acoustic oscillations (BAO) peak. This is especially difficult to model given the very coarse radial bins we have chosen. Due to these limitations of our model, we simply exclude scales larger than $r \gtrsim 100\,\mathrm{cMpc}$ from our analysis.

2.3 Data-model comparison

In the previous Section, we have described how to obtain the two observables of interest (i.e., the QLF and the quasar auto-correlation function) starting from a CLF and a simulation-based analytical description of the halo mass function and of the halo cross-correlation functions. We now provide more details on the actual comparison between our model and observational data.

2.3.1 Overview of observational data

We start by giving a brief description of the data that we compare the model with. Our main goal is to explain the very strong quasar clustering measured by S07 at $z \approx 4$. Thus, we make use of the S07 data for the projected auto-correlation function ($w_\mathrm{p}/r_\mathrm{p}$). Note that the authors assume that the data points are independent (because the quality of the data is not good enough to extract a covariance matrix), so we will do the same and use the S07 errors assuming that the covariance matrix for the data is diagonal. We use the "good fields" data (see S07 for the definition) as they are supposed to be more reliable and – since they show stronger clustering – have proven to be the hardest to reproduce theoretically (e.g., Shankar et al. 2010b). As already mentioned, we exclude the data at very large scales ($r > 100\,\mathrm{cMpc}$) from our analysis because they are particularly challenging to measure both in observations (e.g., Eftekharzadeh et al. 2015) and in simulations (see the end of the last Section).

In the subsequent analysis, we are also interested in reproducing the quasar clustering at lower redshift. For this purpose, we use data from the Baryon Oscillation Spectroscopic Survey (BOSS, Eftekharzadeh et al. 2015; hereafter, E15). We focus on the redshift range $z = 2.2 - 2.8$, where the majority of the BOSS quasars reside. We use the data for the projected correlation function, $w_\mathrm{p}(r_\mathrm{p})$, in the radial range $4\,\mathrm{cMpc}\,h^{-1} < r_\mathrm{p} < 25\,\mathrm{cMpc}\,h^{-1}$. In this region, the E15 data are considered more reliable by the authors and an estimate for the error covariance matrix is available.

One of the key points of our analysis is that, while the QLF includes all quasars known in a given redshift bin, the quasar auto-correlation function is usually measured by considering only quasars above a given luminosity threshold L_thr. This is an important point to take into account in our model

(see eq. 2.6-2.10), as the presence of such a threshold may bias the inferred clustering significantly. The flux limit employed for the S07 measurements is $m_i = 20.2$ (where m_i is the apparent magnitude in the i band). In order to convert this to a value of $L_{\rm thr}$, we first convert the apparent magnitude m_i to an absolute magnitude, M_{1450}, using the $K(z)$ correction[5] (see, e.g., Kulkarni et al. 2019 and references therein). We obtain that $m_i = 20.2$ corresponds to $M_{1450} = -25.72$ at $z = 4$. We then convert this value to a bolometric luminosity by applying the bolometric corrections provided by Runnoe et al. (2012a)[6]. We get a value for the S07 luminosity threshold equal to $\log_{10} L_{\rm thr}/{\rm erg\,s^{-1}} = 46.7$.

As for the E15 clustering data at $z \approx 2.5$, the luminosity threshold that we should employ is more subtle. While the authors consider the entirety of the BOSS sample (Ross et al. 2013) for their clustering analysis, they also show that this sample is highly incomplete at low luminosities. This is an issue in the context of our model, as, when setting a threshold $L_{\rm thr}$, we are implicitly assuming that the sample is complete above the threshold. Given that properly modeling completeness in the E15 sample is outside the scope of this work, we set the value of $L_{\rm thr}$ to the 25th percentile of the luminosity distribution of the observed quasars at $z = 2.5$. This value represents a compromise between taking into account part of the highly incomplete sample of faint quasars that are included in the clustering analysis and minimizing the bias that these quasars generate in the predicted clustering. By considering Figure 3 in E15, we set this threshold value to a $M_i(z = 2)$ magnitude of -25.3. Following Lusso et al. (2015), we convert this to $M_{1450} = M_i(z = 2) + 1.28 = -24.02$, and finally to a bolometric threshold of $\log_{10} L_{\rm thr}/{\rm erg\,s^{-1}} = 46.1$.

As for the QLF, there are many different estimates available. For the sake of consistency with the clustering measurements, we choose to employ the UV-bright quasar catalogue compiled by Kulkarni et al. (2019, hereafter K19). These authors provide a homogenised catalogue of 80,000 color-selected AGN from redshift $z = 0$ to 7.5, together with MCMC-based estimates of the QLF at all redshifts. We employ this dataset and select quasars at different redshifts according to our models. For the model at $z = 4$, we set $3.5 < z < 4.5$ (largely consistent with the S07 high-z sample); in this range, the bright end of the QLF is determined by the same SDSS quasars that are used to compute the clustering (Schneider et al. 2010), whereas the low-luminosity quasars are presented in Glikman et al. (2011). The model

[5]The conversion between m_i and M_{1450} can be made using $K_{i,1450}(z)$, which is defined as: $M_{1450}(z) = m_i - 5\log_{10}(d_{\rm L}(z)/{\rm Mpc}) - 25 - K_{i,1450}(z)$, with $d_L(z)$ being the luminosity distance at redshift z. Following Lusso et al. (2015), we set $K_{i,1450}(z = 4) \approx -1.9$.

[6]The bolometric correction for $\lambda = 1450$ Å is $\log_{10} L_{\rm iso}/{\rm erg\,s^{-1}} = 4.745 + 0.910\log_{10} \lambda L_\lambda/{\rm erg\,s^{-1}}$. $L_{\rm iso}$ is the bolometric luminosity calculated under the assumption of isotropy, and it is related to the real bolometric luminosity L via a factor that accounts for the viewing angle, $L = 0.75\,L_{\rm iso}$

at $z = 2.5$, instead, is entirely determined by quasars observed by the BOSS survey (Ross et al. 2013).

Quasars in the K19 dataset are binned according to their M_{1450} magnitude, and the uncertainties are computed using Poisson statistics. However, as also discussed in K19, the QLF data always present significant systematic errors due to, e.g., uncertainties in the quasar selection. This implies that the quoted uncertainties on the QLF data may be significantly underestimated, as is also evident from the large scatter (up to ≈ 1 dex) between different estimates of the QLF that are available in the literature (e.g., Shen et al. 2020; Grazian et al. 2023). These issues are particularly problematic in our framework, as our goal is to perform statistical inference by simultaneously matching the quasar luminosity and auto-correlation functions, and that can only be done properly if the associated uncertainties are well understood and treated. Therefore, in order to avoid biases in our inference analysis owing to very small formal statistical uncertainties on the QLF, we add a systematic error to every QLF measurement in quadrature to the Poisson ones determined by K19. That is, the uncertainties on our QLF data points are set to be $\sigma^2 = \sigma^2_{\rm sys} + \sigma^2_{\rm count}$, where $\sigma^2_{\rm sys}$ ($\sigma^2_{\rm count}$) stands for the systematic (statistical) uncertainty. We adopt a constant systematic uncertainty of 0.2 dex for the $z \approx 4$ dataset and of 0.05 dex for the $z \approx 2.5$ one. This implies a systematic relative uncertainty of $\approx 45\%$ ($\approx 10\%$) for $z = 4$ ($z = 2.5$). These values are chosen to be similar to the average relative statistical uncertainties at the two redshifts considered ($\approx 40\%$ and $\approx 5\%$ at $z = 4$ and $z = 2.5$, respectively).

As the final step, we convert the values of the quasars' absolute magnitudes in K19, M_{1450}, to bolometric luminosities using the Runnoe et al. (2012a) bolometric corrections. We stress the fact that our results are independent of the adopted bolometric corrections, as our model could easily be expressed in terms of quasars' UV magnitudes only. However, as discussed in Sec. 2.2.1, we choose to convert everything to bolometric luminosities for consistency with previous work on the subject.

2.3.2 Likelihood functions

We employ a Bayesian framework to write the posterior distributions for our model parameters. As described in Sec. 2.2.1, the model consists of a log-normal CLF centered on a power-law dependence of the quasar luminosity on halo mass. The free parameters are the normalization and slope of the quasar luminosity-halo mass relation ($L_{\rm ref}$ and γ, respectively), the logarithmic scatter around this relation (σ), and the fraction of quasars that are active at any given moment ($f_{\rm on}$). The final set of parameters, Θ, is then: $(\sigma, L_{\rm ref}, \gamma, f_{\rm on})$.

We set flat priors on σ and γ, and flat priors on the logarithm of $L_{\rm ref}$ and $f_{\rm on}$ (see e.g. Jeffreys 1946). Our priors span the following parameter

ranges: $\sigma \in (0.1\,\mathrm{dex}, 1.5\,\mathrm{dex})$; $\log_{10} L_{\rm ref}/\mathrm{erg\,s^{-1}} \in (44.0, 46.6)$; $\gamma \in (0.5, 3)$; $\log_{10} f_{\rm on} \in (-3, 0)$. These limits are chosen in order to focus on the region of the parameter space where models are physically motivated (e.g., the scatter in the $L_c - M$ relation is unlikely to be smaller than 0.1 dex).

In what follows, we want to fit the QLF and the auto-correlation function both independently and simultaneously. We can get constraints on these two observables by setting the following likelihood functions:

$$\mathcal{L}^{(k)}(\mathbf{d}^{(k)}\,|\Theta) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})\right), \quad (2.17)$$

where $k \in \{\mathrm{QLF}, \mathrm{corr}\}$ stands either for the quasar luminosity function data or for the auto-correlation function data. As for the other variables, $\mathbf{d}^{(k)}$ stands for the set of n data points with means $\mathbf{y}$ and covariance Σ coming from observations, whereas $\boldsymbol{\mu}$ stands for the set of values predicted by our models.

With the above likelihood, the results for the correlation function ("corr") are found to not be very constraining, as there is a large set of models that produce the correct clustering but substantially under(over)-estimate the number density of bright quasars. Therefore, when quoting results for the correlation function only, we provide an additional integral constraint by imposing that the model matches the observed number density of bright quasars. We integrate the QLF above the luminosity limit used for the clustering measurements (see Sec. 2.3.1), and obtain an estimate for the number density of bright quasars, $n_{\rm bright}$. The associated uncertainty, $\sigma_{\rm bright}$, is determined by using different realizations of the QLF fits from K19. Then, we predict the number of quasars with a luminosity above this threshold, $L_{\rm thr}$, based on our model ($n_{\rm model}$), and use the following likelihood:

$$\mathcal{L}^{(\mathrm{corr+nden})} = \frac{\exp^{-(n_{\rm bright}-n_{\rm model})^2/\sigma^2_{\rm bright}}}{\sqrt{2\pi}\sigma_{\rm bright}}\,\mathcal{L}^{(\mathrm{corr})}. \quad (2.18)$$

Note that we do not fit to the shape of the QLF, but only to the total abundance of quasars above $L_{\rm thr}$. This is an integral constraint that favors models producing a physically reasonable total number of bright quasars.

Finally, we provide joint constraints on the parameters by fitting the QLF and the auto-correlation function simultaneously. In other words, we write the joint likelihood distribution as the product of the two likelihoods (we assume that the two measurements are independent, and weigh the two dataset equally):

$$\mathcal{L}^{(\mathrm{joint})} = \mathcal{L}^{(\mathrm{QLF})}\,\mathcal{L}^{(\mathrm{corr})}. \quad (2.19)$$

Note that for the joint likelihood distribution, we consider $\mathcal{L}^{(\mathrm{corr})}$ (rather than $\mathcal{L}^{(\mathrm{corr+nden})}$), as the QLF already provides an implicit constraint on the total abundance of luminous sources.

Table 2.2: Best-fitting parameter values for our model-data comparison at $z = 4$ (see the main text for definitions of the different parameters, as well as eq. 2.17-2.19). "QLF only" refers to the quasar luminosity function data only, "corr+nden" refers to the auto-correlation function data in conjunction with the number density of bright quasars, and "joint" refers to the combined QLF+auto-correlation function data. The last column shows the minimum value of the normalized chi-squared (see text for details).

Quantity	σ	$\log_{10} L_{\rm ref}$ [erg s^{-1}]	γ	$f_{\rm on}$ [%]	$\chi^2_{\rm norm}$
QLF only	0.38	46.4	0.78	0.2	2.2/5
corr+nden	0.10	44.4	2.99	100	4.6/4
joint	0.11	45.1	2.07	66	12.9/12

2.4 Results

In this section, we describe the results we obtain by fitting our model to the observed quasar luminosity and auto-correlation functions, both independently ("QLF" and "corr+nden" cases) and simultaneously ("joint" case; see Sec. 2.3.2 for the definitions). Henceforth, we will refer to the "QLF" model as "QLF only" in order to distinguish our model from the QLF itself. We first consider the $z = 4$ case – which is the main focus of this paper – and then discuss the results at lower redshift ($z = 2.5$) as well.

2.4.1 Analysis at $z \approx 4$

As a first step, we are interested to know whether our model can reproduce the two observables. We can answer this question by employing a simple optimization algorithm to find the maximum of the likelihood distributions (or, equivalently, of the posterior distributions) for the three cases of interest: quasar luminosity function only ("QLF only"), correlation function + number density of bright quasars ("corr+nden"), and quasar luminosity and correlation functions together ("joint"). The maxima of the likelihoods represent our best-fitting models, which we can then compare directly with observations (see Section 2.4.1.1 for the results of the full parameter inference).

In Table 2.2, we report these best-fitting parameters for the cases mentioned above. Figure 2.3 shows our model predictions at the maximum likelihood parameter values for the CLF, the HOD, the QHMF, the QLF, and the projected quasar autocorrelation function ($w_{\rm p}/r_{\rm p}$); see Fig. 2.1 for a schematic overview of these quantities.

In the top right panel of Fig. 2.3, we show the conditional luminosity functions, CLF($L|M$), as a function of the quasar luminosity L and the halo mass M. The three cases "QLF only", "corr+nden", and "joint" are shown with different colors (blue, orange, and green, respectively). The associated

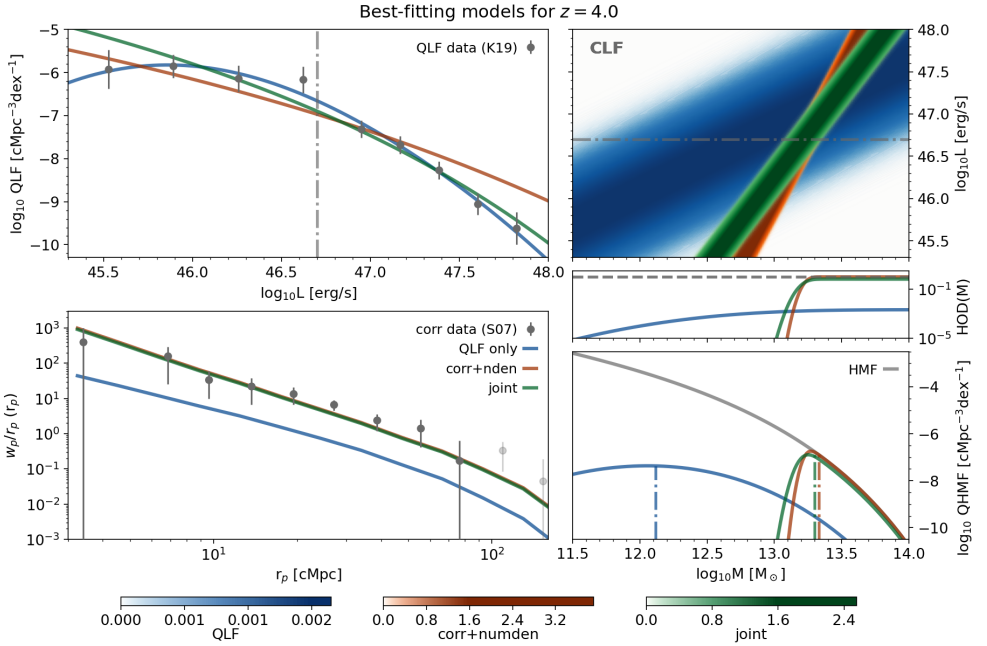

Figure 2.3: Overview of our model-data comparison at $z = 4$. The blue, orange, and green colors refer to the best-fitting models for the "QLF only", "corr+nden", and "joint" likelihood distributions, respectively (see text for details and Table 2.2 for the parameters' values). The upper right panel shows the Conditional Luminosity Function (CLF($L|M$)), with the associated color bars at the bottom representing the probability density. The horizontal dot-dashed gray line in the same panel (and the vertical one in the upper left panel) refers to the luminosity threshold for clustering measurements, $\log_{10} L_{\rm thr}$. Integrating the CLF along the luminosity axis above this threshold gives the Halo Occupation Distribution (HOD; middle right panel), which can be combined with the Halo Mass Function (gray line in the lower right panel), to give the Quasar-Host Mass Function (QHMF; coloured lines in the lower right panel). Vertical dot-dashed lines in the lower right panel are the median values of the QHMF distribution, $M_{\rm med}$ (see eq. 2.12). The QHMF is then used to predict the projected quasar auto-correlation function, $w_{\rm p}(r_{\rm p})$ (lower left panel). The $z \approx 4$, S07 data for the auto-correlation function are also shown in the same panel (data outside the fitting range, $3 < r_{\rm p}/{\rm cMpc} < 100$, are shown as semi-transparent points). The upper left panel shows our predictions for the Quasar Luminosity Function (QLF), together with the $z \approx 4$ data from K19.

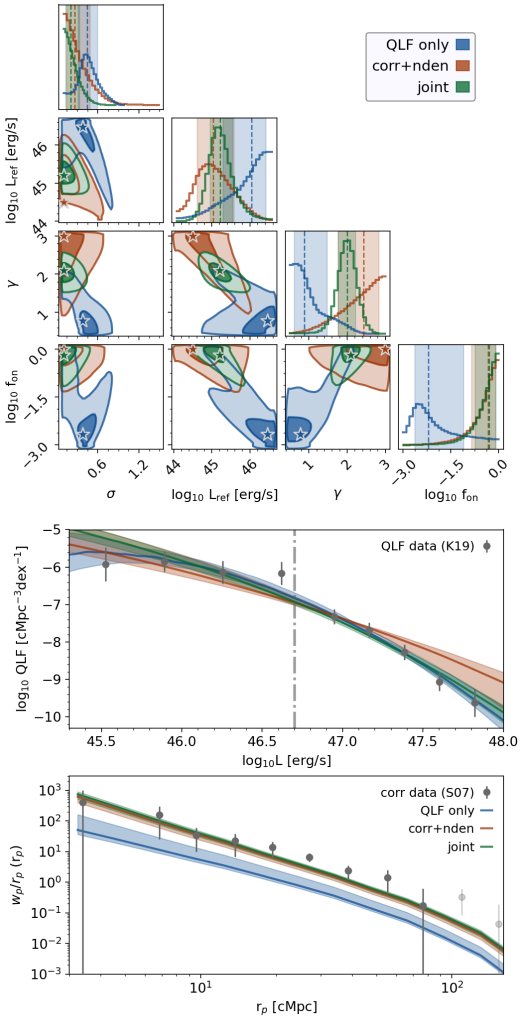

Figure 2.4: *Left:* Corner plots of the 4-d posterior distributions for the different cases described in Sec. 2.4.1 (blue for the "QLF" model, orange for "corr+nden", and green for "joint"). Contours in the 2-d histograms highlight the 1σ and 2σ regions, whereas the dashed lines in the 1-d histograms represent the median values of the parameters (with 1σ errors shown as shaded regions). Best-fitting parameters from Table 2.2 (see also Fig. 2.3) are shown with star symbols in each corner plot. *Right*: Comparison of the predicted quasar luminosity (top) and auto-correlation (bottom) functions with the observational data from K19 and S07, respectively. The color coding is the same as in the left panel. Median values (solid lines) and 1σ uncertainty regions (shaded areas) are obtained by randomly sampling the Markov chains for the posterior distribution 100 times. Data points for the auto-correlation function that are outside of our fitting range (see Sec. 2.3.1) are shown as semi-transparent points in the bottom right panel. The vertical dot-dashed line in the upper right panel is the luminosity threshold for quasar clustering, $L_{\rm thr}$ (see Sec. 2.3).

color bars at the bottom of the Figure represent the probability densities for the different CLF cases. Integrating the CLF above the luminosity threshold $L_{\rm thr}$ (gray dashed-dotted line in the CLF panel), we obtain the halo occupation distribution (HOD; middle right panel; eq. 2.10). Combining the HOD with the halo mass function (HMF), we get the Quasar-Host Mass Function (QHMF; eq. 2.6); this is shown in the bottom right panel, together with the $z = 4$ HMF (gray line). The two left panels show the predictions for the observable quantities: the auto-correlation function is shown on the bottom left, together with data from S07; the quasar luminosity function (eq. 2.2) is shown on top (data are from K19). While the auto-correlation function is obtained from the QHMF via eq. 2.7-2.9, the QLF is the result of integrating along the mass axis of the CLF weighted by the HMF (eq. 2.2).

Overall, looking at the two left panels of Figure 2.3, we conclude that in all cases the models constitute very good fits to the data they are meant to reproduce (see below for the caveat on the "QLF only" case). In order to quantify this, we use reduced chi-squared statistics, $\chi^2_{\rm norm} = \chi^2/\nu_{\rm ndof}$, where $\nu_{\rm ndof}$ is the number of degrees of freedom (i.e., the number of data points minus the number of parameters). We find $\chi^2_{\rm norm} = 2.2/5$, $\chi^2_{\rm norm} = 4.6/4$, and $\chi^2_{\rm norm} = 12.9/12$ for the "QLF only", "corr+nden", and "joint" cases, respectively. These values are also shown in Table 2.2 for reference.

One striking feature of the best-fitting models is that they have very different properties, as can be seen in the top right panel of Fig. 2.3 and the best-fitting parameters shown in Table 2.2. All of them are characterized by low values of the scatter in the quasar luminosity-halo mass relation, σ, but the offset, slope, and normalization of this relation vary significantly between the models.

The "QLF only" model shows an approximately linear relation with a high value of the reference luminosity $L_{\rm ref}$. As a result, the characteristic mass of halos hosting quasars with a luminosity above $L_{\rm thr}$ is low ($\log_{10} M/{\rm M}_\odot \approx 12.35$, lower right panel of Fig. 2.3). This has two consequences. Firstly, halos with $\log_{10} M/{\rm M}_\odot \approx 12 - 12.5$ are much more abundant than the number of observed quasars, and thus a very low active fraction ($f_{\rm on} \approx 0.1\%$) is needed to match the QLF. Secondly, such a low characteristic mass for the halos hosting luminous quasars implies a low value for the quasar auto-correlation function, in conflict with the S07 measurements (lower left panel). In fact, we see that the best-fitting model for the "QLF only" case does not fare well when compared with the clustering data.

The "corr+nden" model, instead, finds a much larger characteristic host mass for bright quasars ($\log_{10} M/{\rm M}_\odot \approx 13 - 13.5$). Such a large mass is achieved by packing quasars in almost all the most massive halos. This is done thanks to a few key ingredients (upper right panel of Fig. 2.3): a low value of the quasar luminosity at the reference mass of $\log_{10} M_{\rm ref}/{\rm M}_\odot = 12.5$, a highly non-linear relation between quasar luminosity and halo mass ($\gamma \approx 3$) and a very low scatter in this relation $\sigma \approx 0.1$. The first two parameters

determine the mass, $\log_{10} \tilde{M}$, at which the quasar luminosity – halo mass relation crosses the luminosity limit $L_{\rm thr}$. The second and third parameters, instead, determine how sharply the HOD drops at masses lower than $\log_{10} \tilde{M}$ (middle right panel of Fig. 2.3). The extreme scenario implied by our best-fitting model is needed to reproduce the measured auto-correlation function. Indeed, the shape and normalization of the S07 data are very well reproduced by our model (Fig. 2.3, lower panel) at the scales considered in the analysis ($3\,{\rm cMpc} \lesssim r_{\rm p} \lesssim 100\,{\rm cMpc}$).

Besides fitting the auto-correlation function, the "corr+nden" model aims to reproduce the number density of quasars above the luminosity threshold $\log_{10} L_{\rm thr}$. This is also achieved by the best-fitting model, which predicts a number density $n_{\rm model} = 3.18 \times 10^{-8}\,{\rm cMpc}^{-3}$, 0.5 standard deviations higher than the observational value of $n_{\rm bright} = 2.73 \times 10^{-8}\,{\rm cMpc}^{-3}$. The shape of the QLF, however, is not well reproduced by the model, because it overpredicts the abundance of very bright systems and underpredicts the abundance of $\log_{10} L/{\rm erg\,s^{-1}} \approx 46 - 47$ quasars. This is due to the fact, despite the very low value of σ, the strong non-linearity in the quasar luminosity-halo mass relation ($\gamma \approx 3$) associates a large fraction of the massive halos to the brightest observable quasars.

When we simultaneously fit both the quasar auto-correlation and the luminosity function ("joint" model), we obtain results that are quite similar to the "corr+nden" case, and are compatible with the same extreme scenario in which quasars are packed in the most massive halos, i.e., a non-linear quasar luminosity-halo mass relation with a steep slope and very small scatter, low value of the quasar luminosity at the reference mass, and a large active fraction of quasars. The quasar luminosity-halo mass relation for the "joint" model is however not as extreme as the one for the "corr+nden" model, as it is characterized by a lower value of the power-law exponent, $\gamma \approx 2$. This has very little impact on the auto-correlation function, as the quasar-host mass functions (lower right panel of Fig. 2.3) are very similar in the two cases. It does have an effect, however, on the shape of the QLF, with the "joint" model providing a better fit, especially at the very bright end.

Overall, the QLF is very well reproduced by the "joint" model, with the exception of the low-luminosity end ($\log L/{\rm erg\,s^{-1}} \approx 45.5$). In this region, the largest differences between the "QLF only" and the "joint" model appear, with the "QLF only" model faring better at predicting a flattening of the shape of the QLF. This flattening, however, is an artificial feature of our model, originating from the prior assumption that halos with a mass lower than $\log_{10} M/{\rm M_\odot} = 11.5$ do not host quasars. We consider this issue not worthy of further investigation, as the faint-end of the QLF is still largely unconstrained by data, and deeper observations are needed to probe its behavior at the high redshift (e.g., Akiyama et al. 2018; Parsa et al. 2018; Giallongo et al. 2019; Harikane et al. 2023; Grazian et al. 2023). Furthermore,

our primary focus here is to interpret the bright quasars that are also probed by clustering surveys. It is possible that a more flexible quasar luminosity-halo mass relation is necessary to account for the abundance of low-luminosity systems.

2.4.1.1 MCMC analysis

Given that our models are a good representation of the observational data, we can proceed further with inference and determine how well the data constrain the model parameters. We explore the posterior distributions using a Markov-Chain Monte Carlo (MCMC) approach. We employ the Python package EMCEE (Foreman-Mackey et al. 2013) to sample the posteriors using the affine-invariant ensemble prescription (Goodman & Weare 2010). We place $m = 48$ walkers distributed randomly in the parameter space and evolve them for $N > 10^5$ steps. We set the final number of steps so that our chains are at least 100 times longer than the auto-correlation time τ (see e.g., Sharma 2017), and thin the chains considering only one element every τ steps in order to account for auto-correlations. We also discard the first 10^3 elements of every chain to account for the burn-in phase.

Figure 2.4 (left panel) shows the corner plot for the 4-d posterior distributions (as a function of $\sigma, L_{\rm ref}, \gamma, f_{\rm on}$) for the three cases considered in the analysis ("QLF", "corr+nden", and "joint"). The best-fitting model for each of these cases, which was discussed above and shown in Fig. 2.3, is highlighted with a star symbol in the corner plots. The samples of the posterior distributions obtained by the Markov Chains are then used to obtain predictions for the quasar luminosity and the auto-correlation functions; we compare these quantities with the data in the right panels of Figure 2.4.

As expected, the "QLF only" and "corr+nden" models peak in very different regions of the parameter space. The "corr+nden" model constrains the parameters to the region with $\sigma \lesssim 0.5$, $\gamma \gtrsim 2$, $\log_{10} L_{\rm ref}/{\rm erg\,s^{-1}} \approx 44.5 - 45.5$, and $f_{\rm on}$ close to unity. This region of the parameter space is the only one that is compatible with the above-mentioned scenario in which bright quasars are active only in the most massive halos. This is also the reason why there are no models predicting stronger quasar clustering than observed (Figure 2.4, right panel), as our models are already predicting the strongest possible clustering compatible with the observed abundance of bright systems.

The "QLF only" model, on the other hand, peaks at lower γ and $f_{\rm on}$, larger $\log_{10} L_{\rm ref}$, and a value of σ which is larger than the "corr+nden" but still moderately low ($\sigma \approx 0.3 - 0.5$). However, the distribution for the "QLF only" case is much more complex, and therefore the resulting constraints on the parameters are not as straightforward. In particular, there is a region of the parameter space that is well within the constraints given by the auto-correlation function, and for which the "QLF only" model also has a

good match with the QLF data (at the $\lesssim 2\sigma$ level). Unsurprisingly, this is the region where the "joint" posterior distribution is located (green contours in Fig. 2.4).

The reason why this region of parameter space can reproduce both the QLF and the auto-correlation function can be understood as follows. As mentioned in the introduction, the behavior of the QLF at the bright end is very different from the one of the HMF, with the latter being characterized by an exponential cutoff that is not present in the QLF. This is usually explained by assuming that the more abundant population of lower-mass halos can also contribute to the population of luminous quasars (see e.g., Ren et al. 2020). Indeed, this is what our fiducial "QLF only" model seems to suggest (see also Fig. 2.3), as the very high quasar luminosity predicted for the $10^{12.5}\,\mathrm{M}_{\odot}$ halo population implies that even with ≈ 0.4 dex of scatter the correct shape of the luminosity function can be reproduced. In this picture, quasars are relatively common phenomena arising in the bulk of the halo population at that redshift, with a very low duty cycle of $\varepsilon_{\rm DC} \approx 0.1\%$. However, even a scenario in which only the most massive halos are active as bright quasars (with a duty cycle $\varepsilon_{\rm DC} \gtrsim 50\%$) can be compatible with the observed shape of the QLF. In this second case, the non-linearity in the quasar luminosity–halo mass relation plays a key role in mapping the exponential cutoff of the HMF into the power-law bright end slope of the QLF, while at the same time packing bright quasars only in the most massive hosts. While both of these scenarios provide a good description of the QLF – differing significantly only at lower luminosities – only the latter is also compatible with a very large clustering length of quasars.

In conclusion, despite the fact that the quasar luminosity and auto-correlation functions alone provide relatively loose constraints on the shape of the Conditional Luminosity Function (CLF), when considered in conjunction they are able to determine a very well-defined region in the parameter space for which a good agreement with all observational data is achieved (right panel of Fig. 2.4). This is the most significant conclusion of our analysis, and we will discuss it further in Sec 2.5.

2.4.2 Comparison with $z \approx 2.5$

Having applied our model to $z \approx 4$ data, it is also important to test whether the model is flexible enough to reproduce observations at lower redshifts, where the observed strength of quasar clustering is not as extreme. We note that our goal in this paper is not to provide a complete and self-consistent evolutionary description of quasar properties across cosmic time, but simply to strengthen the conclusions we have drawn in the previous section by showing that the same framework can also be applied to describe the spatial and luminosity distributions of quasars at different epochs. In particular, we focus on the redshift range $z = 2.2 - 2.8$, where the the BOSS survey (Ross

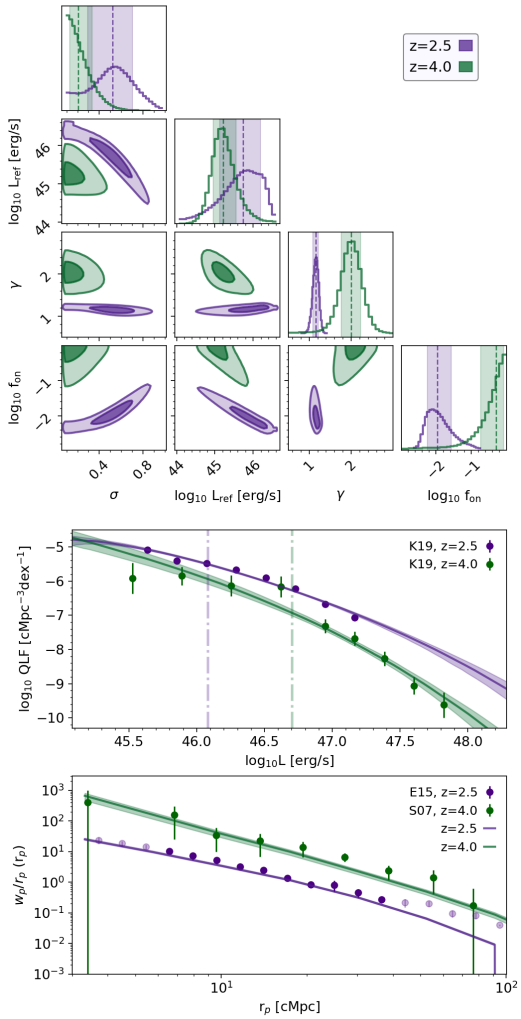

Figure 2.5: Same as Fig. 2.4, but for different redshifts ($z = 4$ in green and $z = 2.5$ in purple). The results always refer to the "joint" model (Sec. 2.4.1). The vertical dot-dashed lines in the upper right panel are the luminosity thresholds, $L_{\rm thr}$, used to measure quasar clustering at the two redshifts. Data points for the auto-correlation function that are outside of our fitting range because they are considered not reliable (see Sec. 2.3.1) are shown with semi-transparent colours in the bottom right panel.

et al. 2013; Eftekharzadeh et al. 2015) has provided solid measurements of the quasar luminosity and auto-correlation functions. We choose this data set because it is sufficiently different from the one at $z \approx 4$ to suggest that the properties of quasars may have varied significantly in a relatively short amount of time.

For simplicity, we focus here on the "joint" models only. In other words, we run the MCMC-based algorithm with the same setup as in Sec. 2.4.1.1 fitting the quasar luminosity function and the auto-correlation function simultaneously. Figure 2.5 shows the resulting corner plots for the posterior distribution of the "joint" model at $z = 2.5$ (purple), together with the one at $z = 4.0$ (green; same as Fig. 2.4) for comparison. We report the values of the resulting 1-d constraints on the model parameters for both redshifts in Table 2.3. In the right panel of Figure 2.5, we show the predictions of our models based on randomly sampling the Markov chains for the posterior distributions, together with the data that we aim to reproduce. We note that, as mentioned in Sec. 2.3.1, we only include the data points for the E15 auto-correlation function at $z \approx 2.5$ in the range 6 cMpc $\lesssim r_{\rm p} \lesssim$ 40 cMpc h^{-1}. This is because data outside this range are not considered reliable and not included in the covariance matrix estimation (see E15). Indeed, we find that our model provides a good match to the E15 data within the fitting range, but it is significantly lower than the measured data at larger scales. Given the strong biases that may be associated with large-scale estimates of the correlation function, we do not consider this issue worthy of further investigation.

The corner plots in Figure 2.5 show that the regions of the parameter space constrained by the two redshifts are quite different. Interestingly, the shape of the $z = 2.5$ posterior distribution exhibits non-trivial behavior in the 2-d projections, yielding tight constraints on the γ parameter, but also strong degeneracies between σ, $\log_{10} L_{\rm ref}$, and $f_{\rm on}$. In general, however, the different parameters are well constrained, even better than at $z = 4$ due to the higher sensitivity of the data. The resulting 1-d posteriors for $z = 2.5$ and $z = 4$ peak at a similar value of $\log_{10} L_{\rm ref}$, but they are quite different for the other parameters. Lower-z results are characterized by a lower value of γ (≈ 1.15) and $f_{\rm on}$ (≈ 0.01), and a higher value of the scatter in the $L - M$ relation, σ.

The top panel of Figure 2.6 shows how these posteriors translate into distributions for the QHMFs (eq. 2.6). In this plot, the QHMFs for $z = 2.5$ and $z = 4$ are shown, together with the HMFs at the same redshifts (semi-transparent lines). Uncertainties on the QHMFs are computed by randomly subsampling the Markov chains for the posterior distributions. The two QHMFs are quite different, reflecting the differences in the level of clustering measured at the two redshifts. In the $z = 4$ case, quasars only reside in the most massive systems ($\log_{10} M/{\rm M}_\odot \gtrsim 13$), with the QHMF distribution tightly following the HMF (see also Fig. 2.3 for the best-fitting model).

At $z = 2.5$, instead, the QHMF distribution has a lower median value ($\log_{10} M/\mathrm{M}_\odot \approx 12.5$) and it is much broader, with a large range of halos of different masses capable of hosting quasars.

The differences in the QHMF translate directly into different measurements for the quasar duty cycle. As discussed in Sec. 2.2.1, we define the quasar duty cycle (eq. 2.13) as the ratio between the QHMF integrated above the median value of its distribution, $M_{\rm med}$ (eq. 2.12), and the HMF integrated above the same threshold. For $z = 2.5$, we find a value of the duty cycle equal to $\varepsilon_{\rm DC} = 0.4 \pm 0.1\,\%$, whereas for $z = 4$ we find $\varepsilon_{\rm DC} = 33^{+34}_{-23}\,\%$. We note that these values are closely related to the values of the $f_{\rm on}$ parameter (Table 2.3), which describes the active fraction of quasars at any given moment. Only in the case of a perfectly deterministic $L - M$ relation (i.e., with zero scatter), however, would we find a duty cycle exactly equal to $f_{\rm on}$. In the presence of scatter in the $L - M$ relation, the shape of the QHMF can vary significantly with respect to the one of the HMF, and this changes the fraction of quasars that are above the threshold luminosity, $L_{\rm thr}$, at any given mass, and hence the quasar duty cycle.

However, we should mention the caveat that these results are obtained by setting two different luminosity thresholds, $L_{\rm thr}$, at the two redshifts considered, according to the minimum luminosities imposed in the respective clustering measurements. As shown in the top right panel of Figure 2.5, the $z = 2.5$ luminosity threshold is ≈ 0.6 dex lower than the one at $z = 4$ ($L_{\rm thr} = 46.1\,\mathrm{erg\,s^{-1}}$ and $L_{\rm thr} = 46.7\,\mathrm{erg\,s^{-1}}$ at $z = 2.5$ and $z = 4$, respectively). Changing the value of $L_{\rm thr}$ may have direct consequences for the QHMF, HOD, and quasar duty cycle, since all these quantities have an explicit dependence on $L_{\rm thr}$ (eq. 2.6-2.13).

In order to provide a fair comparison between these quantities at the two redshifts considered in the analysis, we impose the same $L_{\rm thr}$ at both redshifts by using the $z = 4$ luminosity threshold (i.e., $L_{\rm thr} = 46.7\,\mathrm{erg\,s^{-1}}$) to recompute the above-mentioned quantities at $z = 2.5$. While the duty cycle remains unchanged (within uncertainties), we find that although the QHMF is still very broad, its normalization and median value are lower and higher, respectively. In particular, the median value of the QHMF, $M_{\rm med}$ (eq. 2.12) shifts from $\log_{10} M_{\rm med}/\mathrm{M}_\odot \approx 12.5$ to $\log_{10} M_{\rm med}/\mathrm{M}_\odot \approx 12.8$. This suggests a mild dependence of clustering on luminosity, as more luminous quasars tend to be hosted by more massive halos. However, given that the QHMF distribution is very broad in both cases, there is a strong overlap between the populations of very bright ($\log_{10} L/\mathrm{erg\,s^{-1}} \gtrsim 46.5$) and moderately luminous ($\log_{10} L/\mathrm{erg\,s^{-1}} \approx 46 - 46.5$) quasars in terms of their host halo masses.

We leave a detailed analysis of the implications of our model in terms of the luminosity dependence of quasar clustering for future work. Here, we simply note that even when adopting the same luminosity threshold, we find a remarkable difference between $z = 4$ and $z = 2.5$ quasars. The former are very extreme objects, hosted only by the most massive halos that are

Table 2.3: Constraints on the model parameters based on the corner plots shown in Figure 2.5.

Redshift	σ	$\log_{10} L_{\rm ref}$ [erg s^{-1}]	γ	$f_{\rm on}$ [%]
$z = 2.5$	$0.52^{+0.18}_{-0.22}$	$45.7^{+0.46}_{-0.61}$	$1.15^{+0.06}_{-0.07}$	$1.01^{+1.59}_{-0.04}$
$z = 4$	$0.20^{+0.13}_{-0.08}$	$45.2^{+0.3}_{-0.3}$	$2.00^{+0.22}_{-0.23}$	51^{+32}_{-31}

present at that redshift, representing $4-5\sigma$ peaks in Gaussian random fields (Diemer 2018; see also Sec. 2.1). The latter, instead, are hosted by much more common halos at $z = 2.5$, which are only slightly over-massive with respect to the bulk of the halo population at that redshift ($2-3\sigma$ peaks). For this reason, despite the increase in the quasar number density between $z = 4$ and $z = 2.5$, the quasar duty cycle – which measures how abundant quasars are with respect to their host population – decreases by two orders of magnitudes between the same two redshifts.

In conclusion, our data-model comparison reveals that the same parametrization of the CLF employed at $z = 4$ is also able to reproduce the data at lower-z, with a significant evolution of the CLF parameters reflecting a remarkable change in the physical properties of quasars with cosmic time. In the following, we further discuss the implications of these findings.

2.5 Discussion

In the analysis performed above, we could successfully match the quasar luminosity and auto-correlation functions at two different redshifts provided that: (a) there exists a non-linear relation between quasar luminosity and halo mass, and the non-linearity increases with redshift; (b) the scatter in this relation is fairly small ($\sigma \lesssim 0.3-0.6$) and decreases significantly with redshift; (c) in accordance with this relation, luminous quasars ($\log_{10} L/{\rm erg\,s^{-1}} \gtrsim 46.5$) are hosted by halos with mass $\log_{10} M/{\rm M_\odot} \approx 13-13.5$ ($\log_{10} M/{\rm M_\odot} \approx 12.5-13$) at $z = 4$ ($z = 2.5$); (d) the quasar duty cycle is a strong function of redshift, with a very low $\varepsilon_{\rm DC} \approx 0.4\%$ at low-z that increases to $\varepsilon_{\rm DC} \approx 30\%$ at $z = 4$. In the following, we further elaborate on this picture by investigating its implications for SMBH accretion and growth and by placing it in the context of previous work on the subject. We end the section by highlighting the main strengths and weaknesses of our analysis.

2.5.1 Implications for quasars' physical properties

2.5.1.1 Black hole mass and accretion efficiency

The CLF posits an empirical relation between quasar luminosity and halo mass. However, many quasar population models (e.g., Conroy & White 2013; Veale et al. 2014; Zhang et al. 2023b) built this relation on physical grounds by relating the quasar luminosity to the mass of the central black hole, and this latter mass to the mass of either the host halo or the host galaxy/bulge. We can relate these two approaches by introducing the Eddington ratio η, which is defined by the following relation:

$$L = \zeta \, \eta \, M_{\rm BH}, \tag{2.20}$$

where $M_{\rm BH}$ is the mass of the black hole, and $\zeta = 3.67 \times 10^4 \, {\rm L}_\odot/{\rm M}_\odot$ is a constant factor.

Then, we assume, e.g., that the mass of a black hole is determined solely by the mass of the host halo. In other words, we introduce a probability $P(M_{\rm BH}|M)$ for the mass of the black hole given the halo mass. If we also write the "Eddington ratio distribution" ERDF$(\eta|M_{\rm BH}, M)$ in terms of the other quantities considered, the conditional luminosity function reads:

$$\mathrm{CLF}(L|M) = \int \frac{{\rm d}M_{\rm BH}}{\xi M_{\rm BH}} \, \mathrm{ERDF}\left(\frac{L}{\xi M_{\rm BH}} \middle| M_{\rm BH}, M \right) P(M_{\rm BH}|M). \tag{2.21}$$

In this way, we have related the CLF – which is an empirically determined stochastic relationship between quasar luminosity and halo mass – to two other distribution functions (the ERDF and the black hole mass distribution) that have a clear physical meaning, being related to the physics of black hole accretion and growth.

In order to make this relationship explicit in our analysis, we can simply rewrite the quasar luminosity as the product of the Eddington ratio and the black hole mass (eq. 2.20). In this way, we can explicitly study how these two parameters – albeit completely degenerate – depend on the mass of the host halo, M, according to our model. The middle panel of Figure 2.6 illustrates this dependence. In this panel, we employ the CLF$(L|M)$ relation given by our model to write the probability distribution for the product of the Eddington ratio and the black hole mass-halo mass ratio, $P(\eta M_{\rm BH}/M|M)$. Note that we divide the product $\eta M_{\rm BH} = L/\zeta$ by the halo mass, M, because we expect black hole mass and halo mass to be approximately proportional based on local scaling relations (e.g., Efstathiou & Rees 1988; White et al. 2008; Booth & Schaye 2010; Marasco et al. 2021) and because we can then work with a dimensionless quantity. Redshifts in the middle panel of Figure 2.6 are color-coded as in the top panel and in Figure 2.5. Median values and uncertainties for $P(\eta M_{\rm BH}/M|M)$ are extracted by randomly sampling the

Markov chains for the posterior distributions, as well as the $L - M$ relation given by our model (see the caption for details).

While at $z = 2.5$ the median value of $\eta M_{\rm BH}/M$ shows only a weak trend with halo mass, the situation is much different at $z = 4$, with the product $\eta M_{\rm BH}/M$ strongly correlating with M. This can be achieved by assuming that either the black hole mass, the Eddington ratio, or both increase with halo mass. In other words, for very massive hosts black holes are either particularly massive (with a black hole mass-halo mass ratio higher than for lower-mass counterparts) or efficiently accreting (i.e., with large Eddington ratios). This trend is driven by the fact that the measured strong clustering at $z \approx 4$ requires that the most luminous quasar population is completely dominated by high-mass hosts.

It is also useful to cast these constraints in terms of galaxy stellar masses. This can be done by exploiting one of the parameterizations of the halo mass-stellar mass relation that are available in the literature. Here, we use the redshift-dependent halo mass-stellar mass relation from Behroozi et al. (2013) to rewrite $P(\eta M_{\rm BH}/M|M)$ in terms of the galaxy mass M_*, i.e., $P(\eta M_{\rm BH}/M_*|M)$. For simplicity, we neglect the scatter between stellar mass and halo mass in this conversion. The bottom panel of Figure 2.6 shows how the product of the Eddington ratio and the black hole mass-galaxy mass ratio varies as a function of halo mass. This is especially interesting in light of the fact that there is long-standing evidence in favor of a linear (or quasi-linear) relation between black hole and galaxy masses in the local universe (the so-called $M_{\rm BH} - M_*$ relation, see e.g., Magorrian et al. 1998; Kormendy & Ho 2013; Reines & Volonteri 2015). In the same panel (Figure 2.6), we plot with a dashed red line the expectation for the product $\eta M_{\rm BH}/M_*$ as a function of M, based on assuming the local $M_{\rm BH} - M_*$ relation as measured by Reines & Volonteri (2015), converting galaxy masses to halo masses according to Behroozi et al. (2013), and setting a fixed Eddington ratio of $\eta = 1$. The scatter around this quantity (red shaded region) only considers the scatter in the $M_{\rm BH} - M_*$ relation as quoted by Reines & Volonteri (2015). Due to the fact that the $M_{\rm BH} - M_*$ relation is almost linear, the product between the Eddington ratio and the black hole mass-galaxy mass ratio is almost independent of halo mass, with an average constant value of $\eta M_{\rm BH}/M_* \approx -3.5$.

Comparing this expectation based on local relations to the predictions of our model for $z = 2.5, 4$, we find that for low halo masses ($\log_{10} M/{\rm M}_\odot \lesssim 12.5$) the predictions tend to agree within the uncertainties (see below for the caveat on extrapolating below $\log_{10} M/{\rm M}_\odot \approx 12$). For larger masses, however, the difference between local predictions and our models becomes quite significant. At $z = 2.5$, there is a mild but significant positive trend of increasing $\eta M_{\rm BH}/M_*$ with M. This trend becomes even steeper and tighter at $z = 4$, with the product $\eta M_{\rm BH}/M_*$ ranging from ≈ -3.5 for $\log_{10} M/{\rm M}_\odot \lesssim 12.5$ to ≈ -2 for $\log_{10} M/{\rm M}_\odot \lesssim 13.5$. The trend can be

interpreted considering that the galaxy formation efficiency at $z = 2-4$ peaks at halo masses around $\log_{10} M_{\rm peak}/{\rm M}_\odot \approx 12 - 12.5$; as a consequence, the stellar mass-halo mass relation flattens at masses higher than $\log_{10} M_{\rm peak}$ (Behroozi et al. 2013, 2019). Our model, on the other hand, does not predict any flattening in the quasar luminosity-halo mass relation for masses $\log_{10} M > \log_{10} M_{\rm peak}$. Hence, the product $\eta M_{\rm BH}/M_*$ becomes a steep function of halo mass for $\log_{10} M/{\rm M}_\odot \gtrsim 12.5$. We believe the absence of a flattening in the quasar luminosity-halo mass relation at the high mass end is not simply a consequence of the chosen parametrization for the CLF. By experimenting with different parameterizations of the CLF, we found that any departure from a steep, non-linear relation between quasar luminosity and halo mass is incompatible with the measured value of the clustering (especially at $z = 4$), as a flattening of this relation would lower the characteristic halo mass of bright-quasar hosts. Therefore, we conclude that while galaxy growth appears to be quenched at the high mass end, even at high redshifts (e.g., Behroozi et al. 2019), this does not seem to be the case for black hole growth, as black holes in very massive halos need to be very massive and/or accreting efficiently. Indeed, observational evidence for an evolution of the $M_{\rm BH} - M_*$ relation has been found repeatedly at high-z (implying over-massive black holes) together with signs of an increase in the median value of the ERDF with redshift (e.g., Vestergaard & Osmer 2009; Wu et al. 2022; Maiolino et al. 2024; Pacucci et al. 2023; Stone et al. 2023; however, see Li et al. 2022; Zhang et al. 2023c for a discussion of selection biases).

We conclude by noting that, in our analysis, the shape of the CLF is actually constrained by data only in a limited range of halo masses. For low halo masses, the corresponding quasar luminosities fall in a range where quasar clustering has never been measured and estimates for the QLF are not available (or highly uncertain). On the other hand, for very high halo masses (and hence very high quasar luminosities), quasars become so rare that estimates for the QLF are once again very uncertain. Moreover, if the quasars are luminous enough to be completely above the luminosity threshold for clustering, then the exact behavior of the luminosity as a function of halo mass becomes irrelevant. Therefore, in all panels of Figure 2.6, we show the regions in halo mass where our constraints on the CLF are based purely on extrapolations as dotted lines. This mainly concerns low halo masses ($\log_{10} M/{\rm M}_\odot \lesssim 12.5$) at $z = 4$, and both very low ($\log_{10} M/{\rm M}_\odot \lesssim 12$) and very high ($\log_{10} M/{\rm M}_\odot \gtrsim 13.5$) masses at $z = 2.5$.

2.5.1.2 Quasar lifetime and the growth of high-z black holes

While the empirical relation between quasar luminosity and halo mass gives valuable information on the connection between black holes, their accretion efficiency, and their host halos/galaxies, another key piece of the puzzle

resides in the inferred values of the quasar duty cycle, $\varepsilon_{\rm DC}$. This quantity is defined as the fraction of halos that are hosting bright quasars at any given time (see Sec. 2.2.1). If quasar activity is a stochastic process, however, the duty cycle is also equal to the total fraction of time in which a black hole is active as a bright quasar during the lifetime of an average host halo. In other words, the duty cycle is an average constraint on the total lifetime of a quasar, $t_{\rm Q}$. Following Martini & Weinberg (2001) (see also Martini 2004, Haiman & Hui 2001), we can simply assume that the characteristic lifetime of a halo is roughly equal to the age of the universe $t_{\rm U}(z)$, and get an estimate for the quasar lifetime by writing $t_{\rm Q} = t_{\rm U}(z)\ \varepsilon_{\rm DC}$. Using the values of $\varepsilon_{\rm DC}$ obtained in Sec. 2.4.2, we get $t_{\rm Q} \approx 0.1 - 1$ Gyr at $z \approx 4$, and $t_{\rm Q} \approx 10 - 15$ Myr at $z \approx 2.5$.

The quasar lifetime is one of the most fundamental quantities for understanding the role that SMBHs play in a cosmological context. According to the standard picture of SMBH growth (e.g., Lynden-Bell 1969), luminous quasars are powered by gas accretion onto a SMBH, and the rest mass energy of this material is divided between the small fraction ($\approx 10\%$) of radiation that we observe, and the growth of the black hole. In this picture, a phase of luminous quasar activity translates directly into a buildup of mass for the central SMBH. This provides a direct connection between the total luminosity emitted by quasars over cosmic time and the total mass residing in SMBHs in the local Universe (the so-called "Soltan argument", Soltan 1982). If the quasar lifetime is long compared to the Hubble timescale (i.e., the duty cycle is large), then the buildup of the total SMBH mass has taken place in only a small fraction of host galaxies that were active as bright quasars for a large fraction of their lifetimes. A short quasar lifetime, on the other hand, implies that most galaxies have undergone a brief bright quasar phase during their evolution history. The results of our analysis suggest that the latter scenario is valid at cosmic noon ($z \approx 1 - 3$), when most of the SMBH growth has taken place (e.g., Shen et al. 2020). The short quasar lifetime we find at $z \approx 2.5$ is, in fact, a direct consequence of the fact that quasar activity at that redshift takes place in relatively common halos with a broad distribution of host halo masses (top panel of Figure 2.6). Opposite conclusions can be obtained by considering our $z \approx 4$ results. In this case, we find that the large duty cycle translates into a quasar lifetime that is a large fraction of the Hubble time ($t_{\rm Q} \approx 0.1 - 1$ Gyr). This implies that SMBH growth may be radically different in the young Universe as compared to cosmic noon. As suggested by the $z \approx 4$ QHMF in Figure 2.6 (top panel), quasar activity at high z takes place only in the few most massive halos that are present at that redshift, and hence these systems are active as bright quasars for a large fraction of cosmic time.

Estimating the quasar lifetime at high z is even more compelling in light of the fact that observations of very massive black holes powering luminous quasars at $z \gtrsim 5$ challenge our standard paradigm for black hole

formation and growth (Mazzucchelli et al. 2017b; Farina et al. 2022). In the standard picture, black holes follow an Eddington-limited exponential growth with a timescale that is equal to the "Salpeter time", $t_S \approx 40$ Myr. At high z, models suggest that there is just enough cosmic time to grow the observed SMBH masses starting from massive seeds of $\approx 10^3 - 10^5\,M_\odot$ (e.g., Inayoshi et al. 2020). For this reason, gauging the quasar lifetime is important because it offers an indirect probe of whether sustained accretion on SMBHs can take place at high z in the form of bright quasar activity. The long lifetime we infer at $z \approx 4$ is indeed consistent with this picture, providing an argument in support of models for Eddington-limited growth of high-z black holes. We can provide a rough estimate for this argument by considering as a characteristic host halo mass the median value of the $z \approx 4$ QHMF (Figure 2.6, top panel), $\log_{10} M_{\rm med}/M_\odot \approx 13.3$. If we assume accretion at the Eddington rate ($\eta = 1$), we can translate this characteristic halo mass into a black hole mass using the relation between $\eta\, M_{\rm BH}/M$ and M (middle panel of Figure 2.6): we get $\log_{10} M_{\rm BH}/M_\odot \approx 9$ (Kollmeier et al. 2006). By assuming a seed mass of $10^2\,M_\odot$ ($10^5\,M_\odot$), we find that a total quasar lifetime of ≈ 600 Myr (≈ 350 Myr) is required to grow the black holes under the assumption of Eddington-limited accretion. This is in good agreement with the estimate for t_Q obtained above[7]. This simple argument shows how studying the demographic properties of quasars (such as their abundance and clustering) can place indirect constraints on the formation and evolution history of SMBHs.

Alternative estimates for the quasar lifetime can be obtained by a number of other methods (for an overview, see Martini 2004). Interestingly, results from studies of the quasar proximity effect at high z (Khrykin et al. 2016, 2019) paint a rather different picture than the one suggested here, finding values of the quasar lifetime that are several orders of magnitudes smaller (see also Davies et al. 2019; Eilers et al. 2021). Khrykin et al. (2021) compiled a set of HeII proximity zone measurements for $z \approx 3-4$ quasars, and inferred a log-normal quasar lifetime distribution with a mean of $t_Q \approx 0.2$ Myr and a standard deviation of ≈ 0.8 dex. It is important to note, however, that proximity zone measurements are sensitive only to a fraction of the past quasar lightcurve (up to ≈ 30 Myr for HeII). Clustering measurements, on the other hand, provide integral constraints on the total lightcurve emitted by quasars over the entire history of the Universe. In other words, they are

[7]This estimate assumes that black holes grow at the Eddington limit for their entire history. The demographic properties of quasars at the present time, however, do not constrain black hole growth on a timescale larger than the inferred value of t_Q. We can provide an alternative argument to link the quasar duty cycle to the growth of black hole mass by considering the characteristic luminosity of our quasar sample $L \gtrsim L_{\rm thr}$, and convert that to an accreted black hole mass by assuming a radiative efficiency of $\approx 10\%$ and a total lifetime t_Q. We get $\log_{10} M_{\rm BH}/M_\odot \approx 8.7 - 9.7$ for $t_Q = 0.1 - 1$ Gyr, which again points to the fact that black holes can grow out to very high masses based on our inferred duty cycle.

only sensitive to the zeroth moment of the quasar lightcurve distribution (i.e., the aggregate probability of the lightcurve). The discrepancy between lifetime estimates for proximity zone sizes and clustering measurements, then, may suggest that quasar lightcurves exhibit non-trivial variations on timescales close to the ones probed by proximity zones. With this respect, exploring the full probability distribution associated with quasar lightcurves in the context of our quasar demographic model would provide a way to connect these very different observational probes of quasar activity in a single consistent picture. We will investigate this point in future work.

2.5.2 Comparison with previous work

The model presented in this work builds on a long-standing tradition of interpreting quasar observables via population modeling, i.e., by linking quasars to the well-known population of halos (or sometimes galaxies) according to some empirical/phenomenological prescriptions. Explaining the observed relative abundance of quasars at different luminosities (i.e., the QLF) within such frameworks has been achieved many times, with a large variety of empirical models and physical prescriptions employed (e.g., Efstathiou & Rees 1988; Wyithe & Loeb 2003; Croton 2009; Conroy & White 2013; Fanidakis et al. 2013; Veale et al. 2014; Caplar et al. 2015; Weigel et al. 2017; Ren & Trenti 2021; Zhang et al. 2023b). The bottom line is that the QLF is pretty straightforward to model starting from the hierarchical growth of structures predicted in the ΛCDM framework. On the other hand, the QLF alone does not place tight constraints on key properties of quasars such as their black hole mass, accretion rate, lifetime, and host halo mass, not even in the context of redshift-dependent models (e.g., Wyithe & Padmanabhan 2006; Wyithe & Loeb 2009; Veale et al. 2014). Indeed, our analysis in Sec. 2.4.1 (Fig. 2.4) suggests that a very wide variety of model parameters can be in good agreement with the QLF. As shown by Veale et al. (2014), alternative parametrizations would fare nearly equally well at all redshifts. The large uncertainties on the actual shape and normalization of the QLF that are due to the significant systematics involved in these measurements (Kulkarni et al. 2019) exacerbate this issue, especially at high redshift.

For this reason, considering the independent constraints coming from quasar clustering is extremely useful, as they provide constraints on the masses of the halos that are capable of hosting quasars. Reproducing the clustering of low-redshift ($z \lesssim 2.5$) quasars has been shown to be possible both in empirical models (e.g., Kauffmann & Haehnelt 2002; Hopkins et al. 2007b; Croton 2009; Shankar et al. 2010a; Conroy & White 2013; Aversa et al. 2015; Shankar et al. 2020), semi-analytic models (e.g., Bonoli et al. 2009; Fanidakis et al. 2013; Oogi et al. 2016) and cosmological hydrodynamical simulations (e.g., DeGraf & Sijacki 2017). All of these studies, however, show a significant tension with the clustering measurements at redshift $z \gtrsim 3$.

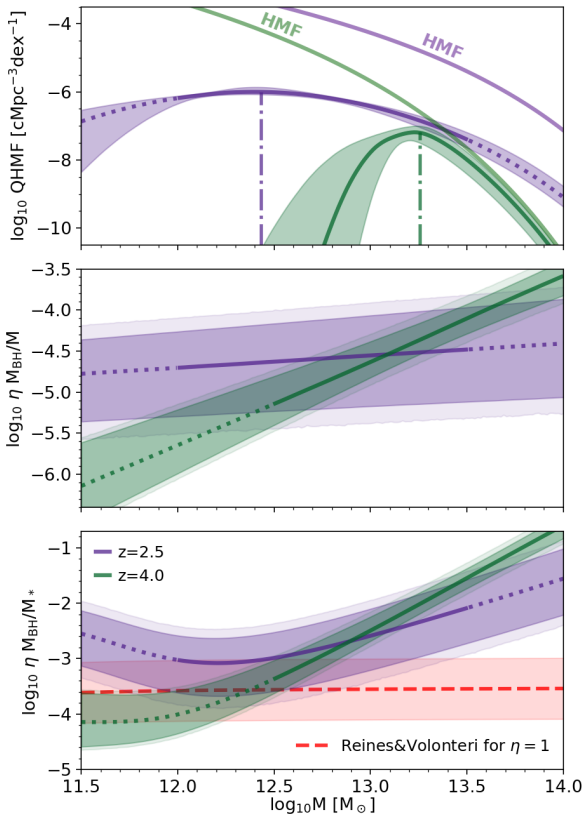

Figure 2.6: *Top:* Quasar-host mass function (QHMF) at $z = 2.5$ (solid purple line) and $z = 4$ (green), according to our model. These functions and their respective uncertainties are the median and the 16th and 84th percentiles of the distributions obtained by randomly subsampling the Markov chains of the posteriors shown in Figure 2.5. The halo mass functions (HMFs) for both redshifts are shown with semi-transparent lines, whereas the dashed-dotted lines indicate the median values of the QHMF distributions. In all panels, regions in the halo mass spectrum where the behavior of the conditional luminosity function (CLF) is purely extrapolated and not explicitly constrained by data are shown with dotted lines. *Middle:* Same as the top panel, but showing the dependence on halo mass of the product between the Eddington ratio (η) and the black hole-halo mass ratio ($M_{\rm BH}/M$). In this case, there are two sources of scatter: the uncertainty on the model given by the posterior distribution and the intrinsic scatter coming from the σ parameter in the CLF. We plot the former with a darker shading, whereas the total contribution of the two sources of scatter is shown with a lighter shading. *Bottom:* Same as the middle panel, but showing the quantity $\eta\, M_{\rm BH}/M_*$ instead (with M_* being the galaxy mass). The relation between halo mass and galaxy mass is taken from Behroozi et al. (2013). The red dashed line shows the prediction for $\eta\, M_{\rm BH}/M_*$ assuming the Reines & Volonteri (2015) relation between black hole and galaxy masses (with the shading showing the scatter in the relation), and setting $\eta = 1$.

The implications for the strong clustering measured by S07 at $z \approx 3-4$, in particular, have been discussed by White et al. 2008, Wyithe & Loeb 2009, and Shankar et al. 2010b. White et al. (2008) assume that the quasar luminosity-halo mass relation is linear, and find that the total number density of bright quasars can be reconciled with the linear bias measured by S07 only if the scatter about this linear relation is very small ($\sigma \lesssim 0.3$ dex). They also find that their conclusions are strongly dependent on the specific functional form assumed for the linear bias-halo mass relation $b(M)$: the Mo & White bias (Mo & White 1996; Jing 1998) is found to be marginally compatible with data, whereas the Sheth, Mo, & Tormen one (Sheth et al. 2001) is inconsistent with the measured bias at the $\approx 2\sigma$ level. Adopting the Sheth, Mo, & Tormen functional form of $b(M)$ after showing that it is a better fit to N-body simulations, Shankar et al. (2010b) interpret the S07 data in the context of an evolutionary model for supermassive black holes, and they strengthen the conclusion that there is tension between the measured bias and the theoretical predictions at $z = 4$. Similar results are found by Wyithe & Loeb (2009), who advocate for a contribution of a merger-driven bias to the $z = 4$ clustering (see also Bonoli et al. 2010; Cen & Safarzadeh 2015 for a discussion of the impact of assembly bias on quasar clustering).

Our work shares some similarities with the three studies mentioned above: we also assume a direct relation between quasar luminosity and halo mass and use the quasar clustering data to infer the specific shape of this relation. Key conclusions of our analysis can also be found in these former attempts to explain the S07 observations: in Sec. 2.4.2, we find that bright quasars need to be hosted by very massive ($\log_{10} M/\mathrm{M}_\odot \gtrsim 13$) halos, and, as a consequence, the quasar duty cycle is a significant fraction of unity ($\varepsilon_{\mathrm{DC}} \approx 10-100\%$). In agreement with White et al. (2008) and Shankar et al. (2010b), we conclude that a relatively small scatter ($\sigma \lesssim 0.3$; Table 2.3) in the quasar luminosity-halo mass relation is necessary to explain the S07 measurement. As also done by Wyithe & Loeb (2009), we adopt a more flexible parametrization of this relation by assuming that it can be non-linear, and find that a steep slope ($\gamma \gtrsim 2$) achieves a much better fit to the data.

The major novelty that our work brings to the understanding of this problem, however, does not reside in the interpretation of the results, but rather in the framework we use to build our model. As explained in Sec. 2.2, we extract the correlation function and the relative abundance of quasars directly from extremely large-volume cosmological N-body simulations, using a novel method to quickly compute the quasar auto-correlation function for any quasar-host mass distribution. In this way, we can directly compare our predictions for the quasar projected correlation function with the S07 observational data. Our model – being based on N-body simulations – naturally accounts for the non-linear contributions to quasar clustering that are essential to interpret the S07 clustering measurements correctly,

especially at scales $r \lesssim 10-15\,\mathrm{cMpc}$. Using this approach, we thus achieve a much more solid data-model comparison, as we do not have to resort to the notion of large-scale linear bias, which is significantly uncertain for strongly biased systems (Diemer 2018) and which discards the information about the shape and the physical features that lie in the S07 data points.

Indeed, according to the statistical analysis performed in Sec. 2.4.1, we find our model can match the data with a satisfactory level of accuracy (i.e., reduced chi-squared ≈ 1), suggesting that, despite being rather extreme, the S07 data can be explained in the context of the standard framework in which clustering of dark matter halos is solely dictated by their mass, without the need to invoke any contributions from assembly/merger bias. We note that we cannot exclude, of course, that such a contribution is present. If that is the case, it would imply that the mass function of $z = 4$ bright-quasar hosts may be somewhat less skewed towards very large halo masses ($\log_{10} M/\mathrm{M}_\odot \gtrsim 13$). We leave an assessment of the role that merger bias plays in cosmological simulations to future work.

Finally, we note that several measurements of the characteristic mass of quasar-hosting halos at $z = 2-4$ are available in the literature. They employ quasar-quasar (S07; E15; Timlin et al. 2018) and quasar-galaxy (Trainor & Steidel 2012; Ikeda et al. 2015; García-Vergara et al. 2017; He et al. 2018; García-Vergara et al. 2019) clustering, as well as gas kinematics in the circumgalactic medium (CGM) of quasar-hosting galaxies (Fossati et al. 2021; de Beer et al. 2023). While the host halo masses predicted by these studies vary significantly, in the present work we have decided to focus on the S07 and E15 measurements of SDSS/BOSS quasars only, because these quasar samples are entirely spectroscopic and thus free from any low-redshift contaminants. However, the same analysis described in this paper could also be performed by taking into account the other clustering measurements mentioned above.

2.5.3 Caveats and final remarks

As shown schematically in Fig. 2.1, the results presented in this work depend on two key ingredients: the choice of the CLF, and the extraction of the halo mass function and the halo (cross-)correlation functions from cosmological N-body simulations. In the following, we will discuss the strengths and weaknesses of our method by considering these components in turn. Let us start with the latter: there are multiple sources of uncertainty in the final estimates we obtain for the halo mass function and the halo cross-correlation functions. First, despite the fact that the box sizes of the simulations employed here are among the largest ever run (Angulo & Hahn 2022), halos are so rare at the very massive end ($4-7\sigma$ peaks in the density field) that the results of simulations at these masses suffer from significant noise. In order to circumvent this issue – and to extrapolate the results of simulations

to the highest mass possible – we used analytical functions to fit the data extracted from the simulations (Sec. 2.2.2). These analytical fits, however, are not perfectly accurate and contribute some systematic errors to our final model predictions.

Nonetheless, we believe that these sources of error can be neglected in our data-model comparison (Sec. 2.4). This is because – as also noted in Sec. 2.2.2 – the observables we are trying to reproduce, i.e., the QLF and the projected correlation function, suffer from significant statistical and systematic uncertainties (as high as $\approx 100\%$ at $z = 4$ and $\approx 30\%$ at $z = 2.5$). In Sec. 2.2.2.1-2.2.2.2 and Appendix 2.B, we assess how well our fitting functions reproduce simulations, and show that their relative accuracy is generally $\lesssim 5 - 10\%$ for both the halo mass function and the cross-correlation functions. For very small ($r \lesssim 5\,\mathrm{cMpc}$) and very large ($r \gtrsim 100\,\mathrm{cMpc}$) scales, measuring the cross-correlation functions in simulations is particularly challenging, especially at the high mass end. The small-scale behavior is highly affected by halo exclusion effects, whereas at large scales the finite size of the simulated boxes reduces the number of pairs, and the baryon acoustic oscillation (BAO) peak makes the shape of correlation functions difficult to capture with our coarse radial bins. As a consequence, our fitting functions are also subject to larger errors at both of these scales. However, these errors do not have a significant impact on our final results, as observational data are also very uncertain at the same scales; for this very same reason, we have excluded the S07 measurements at very large scales ($r > 100\,\mathrm{cMpc}$) from our analysis (see Sec. 2.3.1).

As for the extrapolation of cross-correlations functions to masses higher than the ones that we can probe with our simulations, we have argued that such extrapolation is well motivated by considering the case of $z = 2.5$ in Appendix 2.C. Furthermore, we note that the accuracy of this extrapolation does not have a significant impact on our results: this can be determined by looking at the QHMFs in Figure 2.6 (top panel). At both redshifts, the quasars hosted by halos whose mass is not well represented in simulations are only a small fraction of the total number of quasars (e.g., $\lesssim 5\%$ at $z = 4$). This implies that their actual contribution to the quasar auto-correlation function is negligible compared to the uncertainty in the data.

Other possible sources of uncertainty in our model predictions that we have not discussed yet are the cosmology assumed in the simulations and the exclusion of sub-halos in the creation of the halo cross-correlation functions. Cosmological parameters such as σ_8 and Ω_m are predicted to have a significant effect on the collapse of structures in the standard ΛCDM model, and consequently on the spatial distribution of very massive halos at all reshifts. Studying the impact of these parameters on our final predictions for the clustering of quasars is beyond the scope of this work. Given the current large relative uncertainty on the data, however, we believe that including

variations of the cosmological parameters in our inference procedure would have little effect on our final results.

The exclusion of sub-halos is motivated by the fact that we are only considering clustering measurements at medium-large scales, which are not affected by the distribution of quasars inside a single halo – the so-called one-halo term in HOD models (e.g., Cooray & Sheth 2002). In principle, including the contribution of sub-haloes may boost the large-scale clustering too, as having multiple quasars living in the same dark matter halo implies a larger number of large-scale pairs. In practice, however, sub-haloes are less massive than centrals, and thus they do not tend to host very bright quasars according to the CLF found in Sec. 2.4. We have included sub-haloes in some test runs and verified that large-scale clustering changes only at the percent level, and significant differences are only present at $r \lesssim 1-2\,\mathrm{cMpc}$ even for the most massive halos. On top of that, we note that all of the effects discussed here go in the direction of an enhancement of the predicted clustering, and do not affect the main conclusion of this paper, i.e., that the very strong clustering measured at $z = 4$ can be reproduced with standard assumptions of bright quasars inhabiting massive halos.

In this work, we have assumed one, very simple parametrization for the CLF. We believe that this simple framework is a strength of our model, as it provides very clear physical insight into the formation of quasars and their connection with the hierarchical growth of structures in the context of a ΛCDM universe. On the other hand, we have tested this basic parametrization on a relatively small amount of (very uncertain) data. We have done this on purpose: the main focus of this paper is on reproducing quasar clustering at $z = 4$, and given the quality of the data we have at the present moment, a more sophisticated choice for the CLF would likely have been too flexible to be constrained. However, it is possible that extending our model to a larger/higher signal-to-noise ratio (S/N) dataset, e.g. at $z = 0-2$, would be possible only with a more sophisticated parametrization for the CLF. We leave a thorough examination of different prescriptions for the CLF for future work. In particular, we plan to apply our model to the multiple measurements of quasar clustering available at low-z (e.g., Porciani et al. 2004; Croom et al. 2005; Ross et al. 2009) as well as to the analyses of the dependence of clustering on luminosity at the same redshifts (e.g., Porciani & Norberg 2006; Shen et al. 2009; Eftekharzadeh et al. 2015).

2.6 Summary

We have introduced a novel framework that makes use of multiple cosmological N-body simulations to efficiently predict quasar observables such as the quasar luminosity function (QLF) and the quasar auto-correlation function. The halo mass function and the cross-correlation functions of halos with

different masses are extracted from the dark-matter-only (DMO) versions of the FLAMINGO simulations and used to inform analytical fitting functions. These form the backbone of the model, which is then completed by the choice of a conditional luminosity function (CLF) that links halo masses to quasar luminosities. With these ingredients, we are able to predict the clustering and the luminosity function of quasars, as well as other key properties such as the mass distribution of quasar-hosting halos and the quasar duty cycle (Figure 2.1).

We focus our analysis on the extremely strong clustering measured by Shen et al. (2007) at $z \approx 4$, with the goal of determining whether we can reproduce this measurement in the context of our model. We use a simple parametrization for the CLF, assuming a power-law dependence of quasar luminosity on halo mass ($L \propto M^{\gamma}$) with a log-normal scatter σ. We fit the $z = 4$ QLF and projected correlation function both independently and jointly, in order to gain insight into the best-fitting parameters for each of the cases considered. We then turn our attention to lower-z data, and show that our model can also match the measurements of the same quantities at $z \approx 2.5$ (Ross et al. 2009; Eftekharzadeh et al. 2015), albeit with significantly different values of the model parameters.

We summarise here the main findings of the analysis described above:

- Quasar clustering and abundance measurements at $z \approx 4$ require quasars to reside in the most massive halos at that redshift, with a characteristic mass of $\log_{10} M/\mathrm{M}_{\odot} \gtrsim 13$ (Figure 2.3). This implies that the relation between quasar luminosity and halo mass ($L-M$) is highly non-linear ($\gamma \gtrsim 2$) with a very small amount of scatter ($\sigma \lesssim 0.3$ dex).

- Many different combinations of model parameters can achieve a good fit to the measured QLF at $z \approx 4$ (Figure 2.4). This is because very different empirical prescriptions for the quasar luminosity-halo mass relation (e.g., large scatter and shallow slope or vice-versa) are able to map the exponentially declining end of the halo mass function into the shallower bright end of the QLF. However, the only set of parameters which is also compatible with clustering measurements is the one mentioned above (i.e, a highly non-linear $L - M$ relation with very small scatter), as an increase in the scatter would lower the characteristic mass of quasar-hosting halos, and thus decrease the clustering predicted by our model.

- In order to match the total number density of bright $z \approx 4$ quasars in models in which quasars reside in sufficiently high halo masses to reproduce the observed clustering, the active fraction of quasars (f_{on}) has to be close to unity. This implies that high-z quasars shine for a large fraction of the Hubble time, with a duty cycle in the range $\varepsilon_{\mathrm{DC}} = 10 - 60\%$. In turn, this duty cycle results in a large total quasar

lifetime $t_{\rm Q} \approx 10^8 - 10^9$ yr, consistent with the standard picture of black hole growth in the young universe.

- The steep $z \approx 4$ relation between quasar luminosity and halo mass contrasts with the well-known prediction of a flattening in the stellar mass-halo mass relation at high mass at every epoch (e.g. Behroozi et al. 2013). This implies that in very massive high-z halos – while the star formation may have been quenched already – the supermassive black hole at the center of the galaxy needs to be either over-massive and/or highly accreting. This may have an impact on the shape and normalization of the black hole mass-galaxy mass relation at high redshift (see e.g., Maiolino et al. 2024; Stone et al. 2023).

- Furthermore, the extremely small scatter ($\sigma \approx 0.1 - 0.3$ dex) inferred for the $L - M$ relation at $z \approx 4$ points to some physical processes enforcing a tight relationship between quasars and their dark matter halo hosts. In other words, the relation between black hole mass and stellar and/or halo mass, together with the distribution of Eddington ratios, all conspire to yield a remarkably low scatter.

- The clustering and relative abundance of quasars at lower redshift ($z \approx 2.5$) can be explained by the same parametric relation between quasar luminosity and halo mass. However, the parameters describing this relation show a significant evolution with redshift (Figure 2.5): the slope of the $L - M$ is significantly shallower ($\gamma \approx 1.15$) than at $z \approx 4$, and the scatter larger ($\sigma \approx 0.5$ dex).

- Overall, our comparison between $z \approx 2.5$ and $z \approx 4$ reveals two radically different pictures in terms of the connection between quasars and their host halo population (Figure 2.6). High-z ($z \approx 4$) quasars are hosted by very massive halos, with a very large occupation fraction (i.e., a large fraction of these halos host bright quasars at any given time). At lower redshift ($z \approx 2.5$), instead, quasars reside in halos with a broad range of masses, with the bulk of the population being characterized by relatively common, $\log_{10} M/{\rm M}_\odot \approx 12.5$ mass halos. As a consequence, only a small fraction of low-z quasars are actively shining at any given moment, with a quasar duty cycle of $\varepsilon_{\rm DC} \approx 0.5\%$. These conclusions are consistent with the standard picture of "cosmic downsizing" of quasars and AGN (e.g., Merloni 2004; Scannapieco et al. 2005; Fanidakis et al. 2012), as the bulk of the quasar population is hosted by progressively smaller halos as redshift decreases.

The framework presented here can be readily applied to interpret quasar clustering measurements at all redshifts. In particular, focusing on very high redshift is especially interesting in light of the fact that the large-scale environment of very bright quasars has been proven hard to pinpoint in

the early universe (e.g., Fan et al. 2023). For example, Arita et al. (2023) recently measured the quasar auto-correlation function at $z \approx 6$ for the first time, finding results broadly consistent with the very strong clustering measured at $z \approx 4$. On top of that, several JWST programs such as ASPIRE (Wang et al. 2023) and EIGER (Kashino et al. 2023; Eilers et al. 2023) are starting to deliver measurements of quasar clustering by probing the distribution of line emitters around bright, $z \approx 6-7$ quasars. Connecting the framework presented here to the upcoming quasar-galaxy cross-correlation measurements from JWST will offer a clear and comprehensive picture of the large-scale environments in which the first quasars formed (see also Costa 2024 for an alternative approach).

As suggested by the results obtained in this work, interpreting quasar properties within a consistent framework that takes into account both their demographics and their spatial distribution can give great insight into the relationship between the hierarchical growth of structures in the Universe and the evolution of supermassive black holes over cosmic time.

2.A Appendix: Obtaining the quasar auto-correlation from the halo cross-correlation functions

Let us consider a stochastic process $\mathcal{N}^{(q)}$, describing – in our case of interest – the spatial distribution of quasars. This distribution is discrete: following Peebles (1980), we divide the volume of interest into infinitesimal elements δV_i, and – given the average quasar density $\bar{n}_q$ – we can write the probability of having a quasar in the volume element δV_1 as:

$$\delta P_1 = \langle \mathcal{N}_1^{(q)} \rangle = \bar{n}_q \, \delta V_1. \tag{2.22}$$

Similarly, we define the two-point correlation function, $\xi(r_{12}) \equiv \xi_{12}$, via the probability of having a quasar in the volume element δV_1 and another one in the volume element δV_2:

$$\delta P_{12} = \langle \mathcal{N}_1^{(q)} \mathcal{N}_2^{(q)} \rangle \equiv \bar{n}_q^2 \, \delta V_1 \delta V_2 \, (1 + \xi_{12}) \, . \tag{2.23}$$

We introduce now the continuous number density field $n^{(q)}(\mathbf{x})$, which we define via the expression:

$$\delta P_{12} = \langle \mathcal{N}_1^{(q)} \mathcal{N}_2^{(q)} \rangle \equiv \langle n^{(q)}(\mathbf{x}_1) n^{(q)}(\mathbf{x}_2) \rangle \, \delta V_1 \delta V_2, \tag{2.24}$$

and write this equation in terms of the density contrast field $\delta^{(q)}$ – defined as $n^{(q)}(\mathbf{x}) = \bar{n}_q \, \left(1 + \delta^{(q)}(\mathbf{x})\right)$:

$$\delta P_{12} = \bar{n}_q^2 \, \delta V_1 \delta V_2 \left(1 + \langle \delta^{(q)}(\mathbf{x}_1) \delta^{(q)}(\mathbf{x}_2) \rangle \right) . \tag{2.25}$$

By comparing this to eq. 2.23, we find that the correlation function can also be expressed as:

$$\xi_{12} = \langle \delta^{(q)}(\mathbf{x}_1)\delta^{(q)}(\mathbf{x}_2)\rangle \equiv \langle \delta_1^{(q)}\delta_2^{(q)}\rangle. \tag{2.26}$$

We want to split the different contributions of quasars to the density field $n^{(q)}$ based on the mass of their host halos. We introduce a set of continuous fields $\{n^{(h)}(M_k)\}$, which represent the distributions of halos for different mass bins centered on M_k and with a width ΔM.

We now make the key hypothesis that the distribution of quasars with a host halo mass in the range $[M_k - \Delta M/2, M_k + \Delta M/2]$ is an unbiased tracer of the underlying distribution of halos, $n^{(h)}(M_k)$. In other words, the quasars at a given host halo mass are just undersampling the distribution of halos, and they are thus described by the same stochastic process. This is the case if the presence of a quasar depends solely on the mass of its host. Thus, we can write the quasar distribution, $n^{(q)}$ in terms of the distributions of halos with different masses by simply weighing them by the relative number of quasars at those masses:

$$n^{(q)} = \sum_k p_k\, n^{(h)}(M_k), \tag{2.27}$$

where p_k represents the probability that a quasar has a host-halo mass in the bin M_k. Using the "quasar-host mass function" (QHMF) introduced in Sec. 2.2.1 (eq. 2.6), we can express this probability as ($\bar{n}_{q,k}$ is the average number density of quasars in the host mass bin M_k):

$$p_k = \frac{\bar{n}_{q,k}}{\bar{n}_q} = \frac{n_{\rm QHMF}(M_k)\,\Delta M}{\int_0^\infty n_{\rm QHMF}(M)\,\mathrm{d}M}. \tag{2.28}$$

Introducing the overdensity definitions for the distributions at different masses, $n^{(h)}(M_k) = \bar{n}_{q,k}\,\left(1 + \delta^{(h)}(M_k)\right)$ we can write:

$$\begin{aligned}\langle n_1^{(q)} n_2^{(q)}\rangle &= \sum_j \sum_k p_j p_k \bar{n}_q^2 \left(1 + \langle \delta_1^{(h)}(M_j)\delta_2^{(h)}(M_k)\rangle\right) = \\ &= \bar{n}_q^2 \left(1 + \sum_j \sum_k p_j p_k \langle \delta_1^{(h)}(M_j)\delta_2^{(h)}(M_k)\rangle\right),\end{aligned} \tag{2.29}$$

where we have made use of the fact that $\sum_k p_k = 1$. Introducing the cross-correlation functions for halos of different masses, $\xi_{12}^{(h)}(M_j, M_k) = \langle \delta_1^{(h)}(M_j)\delta_2^{(h)}(M_k)\rangle$, we can express the quasar auto-correlation as:

$$\xi_{12} = \langle \delta_1^{(q)}\delta_2^{(q)}\rangle = \frac{\langle n_1^{(q)} n_2^{(q)}\rangle}{\bar{n}_q^2} - 1 = \sum_{j,k} p_j p_k\, \xi_{12}^{(h)}(M_j, M_k). \tag{2.30}$$

This proves eq. 2.7, which relates the quasar auto-correlation function $\xi(r_{12}) \equiv \xi_{12}$ to the cross-correlation functions of halos with different masses, $\xi_h(M_j, M_k; r_{12}) \equiv \xi_{12}^{(h)}(M_j, M_k)$.

2.B Appendix: Fitting the cross-correlation terms from simulations

In this Section, we provide details on the fitting we perform to the values of the halo cross-correlation functions, $\xi_h(M_j, M_k; r)$, extracted from simulations. As described in Sec. 2.2.2 and 2.2.2.2, we extract these values from two simulations with box sizes equal to $L = 2800\,\mathrm{cMpc}$ and $L = 5600\,\mathrm{cMpc}$, respectively. For each simulation, we consider only halos in a specific range of masses, so that all the mass bins considered are populated by a sufficient number of well-resolved halos. In particular, we set the following ranges (Sec. 2.2.2): $\log_{10} M/\mathrm{M}_\odot = 11.5 - 13.0$ for $L = 2800\,\mathrm{cMpc}$, and $\log_{10} M/\mathrm{M}_\odot = 12.5 - 13.5$ for $L = 5600\,\mathrm{cMpc}$. We choose a bin width of 0.25 in $\log_{10} M$, so that we have 6×6 cross-correlation terms for $L = 2800\,\mathrm{cMpc}$, and 4×4 cross-correlation terms for $L = 5600\,\mathrm{cMpc}$. Note that the masses M_j and M_k in the expression $\xi_h(M_j, M_k; r)$ do not refer to the center of their respective bins, but rather to the median value of the halo mass function in those bins.

Our goal is then to find a single analytical description of these cross-correlation functions that can represent the two simulations simultaneously. In order to do that, we first divide all the cross-correlation functions $\xi_h(M_j, M_k; r)$ by a reference correlation $\xi_{\mathrm{ref}}(r)$; in formulae, we define $\rho(M_j, M_k; r)$ to be:

$$\rho(M_j, M_k; r) = \xi_h(M_j, M_k; r)/\xi_{\mathrm{ref}}(r). \tag{2.31}$$

In this way, we hope that $\rho(M_j, M_k; r)$ will be only marginally dependent on the scale r. We set $\xi_{\mathrm{ref}}(r) \equiv \xi_h(\tilde{M}, \tilde{M}; r)$, with $\tilde{M}$ representing the $\log_{10} M = 12.5 - 12.75$ bin. This choice is arbitrary, but it is made to ensure that the mass bin sits in the overlap between the mass ranges of the two different simulations we use. We also attempt to minimize any dependences of the cross-correlation functions on cosmology and redshift by expressing all the masses in terms of peak heights $\nu(M)$ (see also Sec. 2.2.2.1).

Finally, we fit a 3-d polynomial $\rho_{\mathrm{fit}}(\nu_j, \nu_k, r)$ to the values extracted from the simulations. We empirically find that a second-degree polynomial in mass and third-degree in the radial dimension fits the data points well enough and at the same time attains a smooth behavior with respect to all three variables (i.e., we prevent overfitting). The errors on the data points are assigned based on the Poisson statistics of the pair counts (eq. 2.16). As also done in Sec. 2.2.2.1, we weigh the errors associated with the two

simulations differently in order to achieve a better fit. In particular, we double the values of the Poisson errors for the $L = 2800\,\mathrm{cMpc}$ simulation, and we halve the ones associated with the $L = 5600\,\mathrm{cMpc}$ box.

Figures 2.7–2.8 show the results of the fitting for the two redshifts considered in this work: $z = 4$ (Fig. 2.7) and $z = 2.5$ (Fig. 2.8). The first row of each plot displays the resulting fitting function $\rho_{\rm fit}(\nu(M_j), \nu(M_k), r)$. Each panel in this row shows the values of $\rho_{\rm fit}(\nu(M_j), \nu(M_k), \bar{r})$ as a function of the two masses M_j and M_k at a different scale $\bar{r}$. The second and third rows show the relative differences $(\rho/\rho_{\rm fit} - 1)$ between our fit and the two simulations considered. The mass ranges that are selected in each simulation are shown as grey boxes in the 2-d mass planes. According to these figures, our simple analytical framework can describe the behavior of cross-correlation functions in a wide mass range with a good degree of accuracy ($\lesssim 5 - 10\%$). Notable exceptions to this can be found for very high masses ($\log_{10} M/\mathrm{M}_\odot \gtrsim 13.2 - 13.3$) and very large scales ($r \gtrsim 100\,\mathrm{cMpc}$). However, these are both expected, as both at large masses and large scales correlation functions are difficult to measure in simulations. Further discussion of this can be found in Sec. 2.2.2.2 and Sec. 2.5.3. Similar conclusions on the quality of our fits can be drawn by looking at the right panels of Figure 2.2 and Figure 2.9, where predictions from our fitting functions are compared to auto-correlation functions extracted from the simulations.

2.C Appendix: Halo mass function and correlation functions for redshift $z = 2.5$

In the main text (Sec. 2.2.2.1–2.2.2.2), we discussed our predictions for the halo mass function and the halo (cross-)correlation functions at $z = 4$. We show here the same results for the other redshift that we consider in our analysis, $z = 2.5$. In Figure 2.9 (left panel), we show the halo mass function extracted from the two simulations ($L = 2800\,\mathrm{cMpc}$ and $L = 5600\,\mathrm{cMpc}$ in teal and red, respectively), as well as our fitting function (eq. 2.14-2.15, gray line). The best-fitting parameter values for $z = 2.5$ are: $A = 0.464$; $a = 3.43$; $b = 0.847$; $c = 1.31$. Note that for the fitting we employ the same mass ranges as we used for $z = 4$ (Table 2.1). This choice is clearly sub-optimal, as halos are much more abundant at lower-z, and therefore mass bins with $\log_{10} M/\mathrm{M}_\odot > 13.5$ are well populated even for the smallest box considered ($L = 2800\,\mathrm{cMpc}$). However, we choose to not take into account masses larger than $\log_{10} M/\mathrm{M}_\odot > 13.5$ for the fitting so that we can benchmark how well our fitting function fares if extrapolated to masses larger than this limit. In this way, we can test whether our fitting framework is valid to interpret the behavior of the halo mass function up to masses higher than the ones we can simulate. As shown in Fig. 2.C (left panel), the trend of the halo mass

function at large masses is well described by the extrapolation of our fitting up to $\log_{10} M/\mathrm{M}_\odot \gtrsim 14$; at higher masses, halos become very rare even at $z = 2.5$, and the halo mass function becomes quite noisy and its behavior highly uncertain.

The right panel of Figure 2.9 shows the halo auto-correlation functions for different mass bins obtained both from simulations (colored points) and from our fitting function (gray lines; see Appendix 2.B). We use 8 mass bins ranging from $\log_{10} M/\mathrm{M}_\odot = 11.5$ to $\log_{10} M/\mathrm{M}_\odot = 13.5$ and with a width of 0.25 dex. Lower mass bins correspond in Fig. 2.C to lower values of the correlation functions, and vice-versa. Once again, the mass ranges employed for our fitting are the same for both $z = 2.5$ and $z = 4$, and do not go higher than $\log_{10} M/\mathrm{M}_\odot = 13.5$. This gives us the possibility of testing how the extrapolation of $\xi_{h,\mathrm{fit}}(M_j, M_k; r)$ fares at larger masses. As explained in Sec. 2.2.2.2, ensuring that our theoretical framework can be extended to very high masses ($\log_{10} M/\mathrm{M}_\odot \approx 14$) is quite important, as – especially at $z = 4$ – a significant fraction of quasars are hosted by this population of massive halos that is not well represented in our simulations.

Extrapolations from our fits are shown in Fig. 2.9 (right panel) with dashed lines. We also extract halo auto-correlation functions from the $L = 5600\,\mathrm{cMpc}$ box for the mass bin $\log_{10} M/\mathrm{M}_\odot = 13.5 - 13.75$; we show these values with golden crosses in Fig. 2.9 (right panel). We see that the extrapolation agrees with simulations at the same level as the points that are used for fitting, with the only exceptions being very small ($r \lesssim 5\,\mathrm{cMpc}$) and very large ($r \gtrsim 100\,\mathrm{cMpc}$) scales. Further discussion on the implications of these results can be found in Sec. 2.5.3.

Acknowledgements

We are grateful to the FLAMINGO team for making their dark matter only simulations available. We acknowledge helpful conversations with the ENIGMA group at UC Santa Barbara and Leiden University. EP is grateful to Roi Kugel for helpful discussion, and to Molly Wolfson, Shane Bechtel, and Silvia Onorato for comments on an early draft of this paper. JFH and EP acknowledge support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 885301). This work is partly supported by funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860744 (BiD4BESt). This work used the DiRAC@Durham facility managed by the Institute for Computational Cosmology on behalf of the STFC DiRAC HPC Facility (www.dirac.ac.uk). The equipment was funded by BEIS capital funding via STFC capital grants ST/K00042X/1, ST/P002293/1,

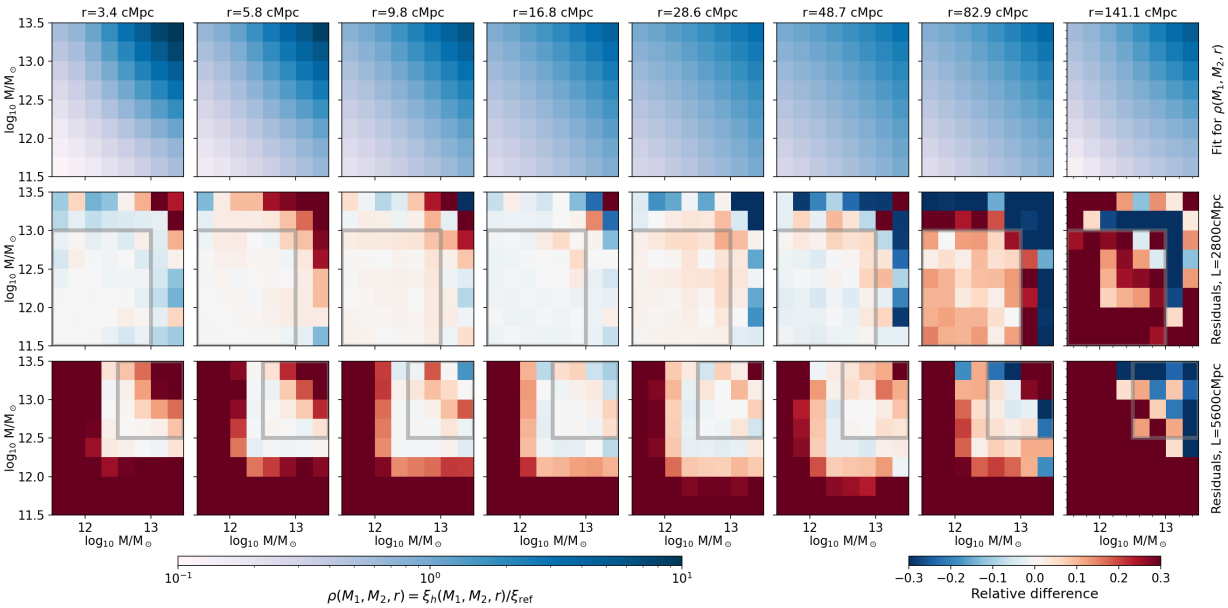

Figure 2.7: Results for the fitting of the cross-correlation terms $\rho(M_1, M_2, r)$ (see Appedix 2.B for definitions). The top row shows the fitting function $\rho_{\rm fit}(M_1, M_2, r)$ as a function of the two masses M_1 and M_2 for different values of the distance r. The second and third rows show the relative difference between the fits and the values measured from the simulations ($L = 2800\,$cMpc and $L = 5600\,$cMpc, respectively). Mass ranges that are adopted for the fitting in each simulation are highlighted by the gray boxes in each mass-mass plane.

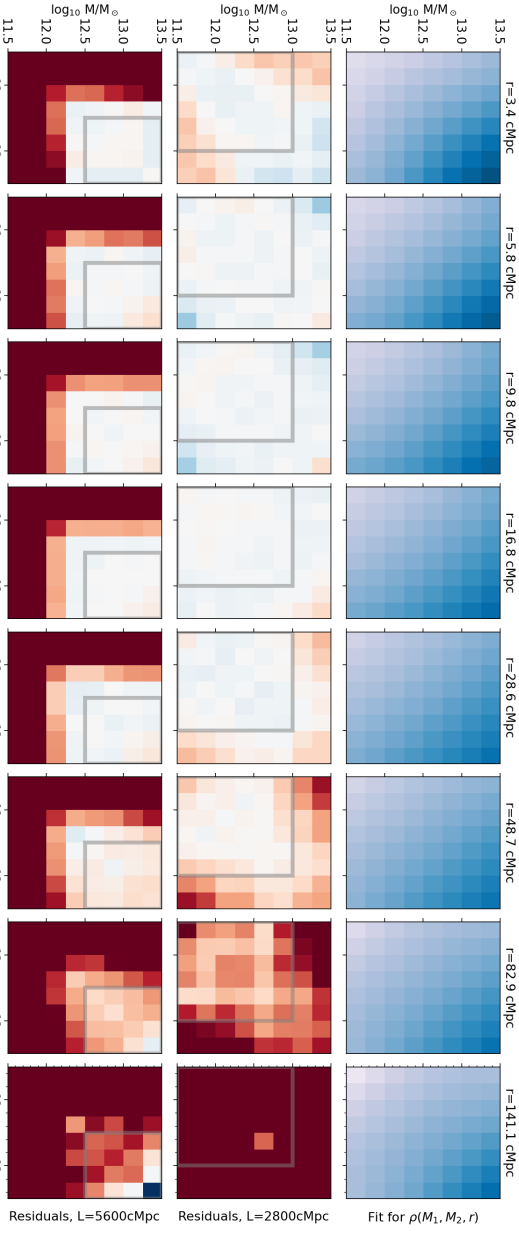

Figure 2.8: Same as Figure 2.7, but for redshift $z = 2.5$.

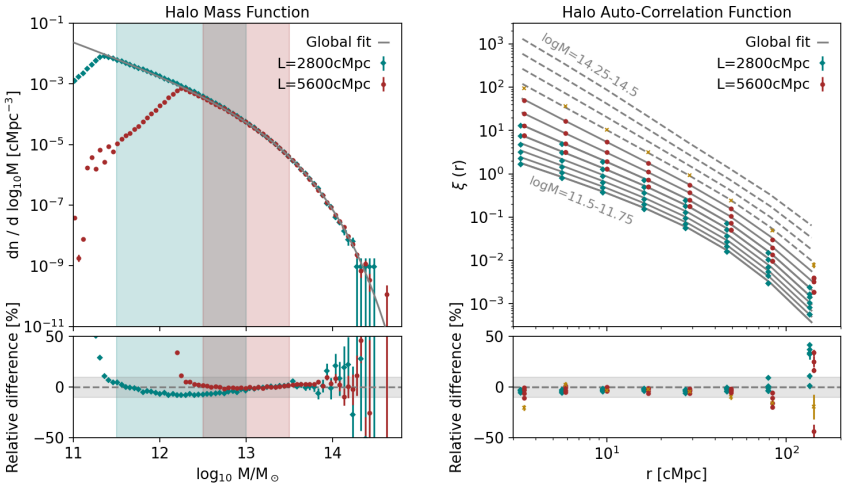

Figure 2.9: Same as Fig. 2.2 but for the snapshots at redshift $z = 2.5$. Golden crosses in the right panel represent the auto-correlation functions measured in the mass bin $\log_{10} M/\mathrm{M}_\odot = 13.5 - 13.75$, in the $L = 5600\,\mathrm{cMpc}$ simulation. This is used as a benchmark to assess how well our fits (dashed grey lines) can be extrapolated to higher masses.

ST/R002371/1 and ST/S002502/1, Durham University and STFC operations grant ST/R000832/1. DiRAC is part of the National e-Infrastructure.

3 A unified model for the clustering of quasars and galaxies at $z \approx 6$

Abstract

Recent observations from the EIGER JWST program have measured for the first time the quasar-galaxy cross-correlation function at $z \approx 6$. The auto-correlation function of faint $z \approx 6$ quasars was also recently estimated. These measurements provide key insights into the properties of quasars and galaxies at high redshift and their relation with the host dark matter halos. In this work, we interpret these data building upon an empirical quasar population model that has been applied successfully to quasar clustering and demographic measurements at $z \approx 2-4$. We make use of a new, large-volume N-body simulation with more than a trillion particles, FLAMINGO-10k, to model quasars and galaxies simultaneously. We successfully reproduce observations of $z \approx 6$ quasars and galaxies (i.e., their clustering properties and luminosity functions), and infer key quantities such as their luminosity-halo mass relation, the mass function of their host halos, and their duty cycle/occupation fraction. Our key findings are: (i) quasars reside on average in $\approx 10^{12.5}\,\mathrm{M}_\odot$ halos (corresponding to $\approx 5\sigma$ fluctuations in the initial conditions of the linear density field), but the distribution of host halo masses is quite broad; (ii) the duty cycle of (UV-bright) quasar activity is relatively low ($\approx 1\%$); (iii) galaxies (that are bright in [O III]) live in much smaller halos ($\approx 10^{10.9}\,\mathrm{M}_\odot$) and have a larger duty cycle (occupation fraction) of $\approx 13\%$. Finally, we focus on the inferred properties of quasars and present a homogeneous analysis of their evolution with redshift. The picture that emerges reveals a strong evolution of the host halo mass and duty cycle of quasars at $z \approx 2-6$, and calls for new investigations of the role of quasar activity across cosmic time.

Published in: **EP**, Joseph F Hennawi, Joop Schaye, Matthieu Schaller, Anna-Christina Eilers, et al. (15 authors), *A unified model for the clustering of quasars and galaxies at $z \approx 6$*, Monthly Notices of the Royal Astronomical Society, Volume 534, Issue 4, November 2024, Pages 3155–3175, doi.org/10.1093/mnras/stae2307
Reprinted here in its entirety.

3.1 Introduction

Supermassive black holes (SMBHs) are thought to be ubiquitous in the Universe, residing at the center of almost every massive galaxy (e.g., Magorrian et al. 1998; Ferrarese & Merritt 2000; Kormendy & Ho 2013). The basic elements of our formation story for these enigmatic objects have hardly changed since their existence was hypothesized, triggered by the discovery of the first quasar (Schmidt 1963). Luminous quasars are powered by accretion onto a SMBH (Salpeter 1964; Zel'dovich & Novikov 1967; Lynden-Bell 1969) and the rest mass energy of this material is divided between the small fraction ($\approx 10\%$) of radiation that we observe, and the growth of the black hole. This implies that the growth of black holes is directly related to the accretion of material powering bright quasars.

But this half-century-old picture is challenged by the existence of luminous high-z quasars powered by $\gtrsim 10^9\,\mathrm{M}_\odot$ SMBHs at $z \gtrsim 6$, well into the epoch of reionization (EoR; Mazzucchelli et al. 2017b; Farina et al. 2022; Fan et al. 2023). Even more puzzling, quasars with similar masses have been discovered at $z \approx 7.5$, merely 700 Myr after the Big Bang (Bañados et al. 2018; Yang et al. 2020, 2021; Wang et al. 2021). The advent of the James Webb Space Telescope (JWST) has made these findings even more compelling, with the record-breaking discoveries of moderately massive SMBHs ($\approx 10^6 - 10^8\,\mathrm{M}_\odot$) at even higher redshift ($z \approx 8 - 11$; e.g., Übler et al. 2023; Maiolino et al. 2024; Kokorev et al. 2023; Larson et al. 2023; Bogdán et al. 2024). How these SMBHs have formed at such early times challenges our understanding of black hole formation and growth. There does not appear to be enough cosmic time to grow them from the $100\,\mathrm{M}_\odot$ seed black holes expected for Pop III stellar remnants (Heger et al. 2003), even if they accrete at the maximal Eddington rate. This has led to an industry of speculation that SMBHs formed from far more massive seeds forming via direct collapse (e.g., Bromm & Loeb 2003) or coalescence of a dense Pop III star cluster (e.g., Omukai et al. 2008).

Addressing this challenge requires integrating SMBH growth into our current picture of galaxy formation and evolution. The tight local scaling relation between SMBHs and galaxy bulges (Magorrian et al. 1998), as well as the need to tap into SMBH accretion as a source of energetic feedback that regulates star formation in massive galaxies (e.g., Benson et al. 2003; Springel et al. 2005; Bower et al. 2006), has led to the modern picture that SMBHs and their host galaxies co-evolve (Bower et al. 2017). In this context, an assortment of cosmological simulation models can produce the massive SMBHs (e.g., Feng et al. 2016; Khandai et al. 2015) that are powering bright high-z quasars starting with massive $\gtrsim 10^4$ $\mathrm{M}_\odot$ seed black holes. These models generically predict that such quasars are hosted by massive ($M_\star \gtrsim 10^{11}$ $\mathrm{M}_\odot$) and highly star-forming (SFR $\gtrsim 100$ $\mathrm{M}_\odot$ yr^{-1}) galaxies,

and reside in the rarest $M \gtrsim 10^{12.5}$ $M_\odot$ halos situated in the most overdense regions of the Universe (Di Matteo et al. 2012; Costa et al. 2014; Feng et al. 2016; Khandai et al. 2015; Barai et al. 2018; Valentini et al. 2021). While these numerical studies establish the plausibility of the existence of high-z quasars, rigorous tests of this theoretical picture have been lacking (Fan et al. 2023; Habouzit et al. 2019).

The key to understanding high-z quasars and SMBH formation in a cosmological context is determining how they are embedded in the evolving cosmic web of dark matter (DM) halos that forms the backbone of all structures in the Universe according to the hierarchical structure formation paradigm. ΛCDM dictates that the clustering of a population of objects, or equivalently the size of the cosmic over-densities that they reside in, is directly related to their host halo masses (e.g. Kaiser 1984; Bardeen et al. 1986; Mo & White 1996). Measuring the masses of the halos that host bright quasars gives precious information not only on the large-scale environment that quasars inhabit, but also – by comparing the observed abundance of quasars with that of the hosting halos – on the fraction of SMBHs that are active as bright quasars at any given time (i.e., the *quasar duty cycle*). In turn, this fraction can be related to the total time SMBHs shine as quasars (or *quasar lifetime*, t_Q; see e.g., Martini & Weinberg 2001; Haiman & Hui 2001; Martini 2004), which is an essential quantity for determining the growth of SMBHs and sets an upper limit to the characteristic timescale of quasar events. For these reasons, a measurement of the clustering of quasars at high redshift is key to unraveling their formation history (e.g., Cole & Kaiser 1989; Efstathiou & Rees 1988).

Quasar clustering studies at lower redshifts are already a fundamental ingredient on which we built our understanding of SMBHs, their accretion mechanisms, and the co-evolution with their host galaxies. Large-sky surveys, such as the Sloan Digital Sky Survey (SDSS, York et al. 2000) and the 2dF QSO redshift survey (2QZ, Croom et al. 2004), have delivered measurements of the auto-correlation function of quasars up to $z \approx 4$ (Porciani et al. 2004; Croom et al. 2005; Porciani & Norberg 2006; Shen et al. 2007; Ross et al. 2009; Eftekharzadeh et al. 2015). These measurements reveal that in the last ten billion years ($z \lesssim 2$), quasars have been tracing halos in a way that is similar to optically selected galaxies, with a linear bias factor close to unity (Croom et al. 2005; Ross et al. 2009). This implies that quasars are hosted, on average, by common, $\approx 10^{12}$ $M_\odot$ halos which, incidentally, are also the ones with the highest star formation efficiency (e.g., Eke et al. 2004; Fanidakis et al. 2010; Fanidakis et al. 2013). At $z \approx 2-4$, however, the clustering of quasars shows a dramatic change from an auto-correlation length, $r_{0,\mathrm{QQ}}$, of $\approx 8\,\mathrm{cMpc}\,h^{-1}$ at $z \approx 2-3$ (White et al. 2012; Eftekharzadeh et al. 2015) up to $\approx 24\,\mathrm{cMpc}\,h^{-1}$ at $z \approx 4$ (Shen et al. 2007). This rapid evolution in quasar clustering implies that quasars live in more massive halos as redshift increases, with a duty cycle that becomes larger as the number of host halos

drops rapidly according to the exponential decline of the halo mass function (Press & Schechter 1974). At $z \approx 4$, the situation seems to be particularly extreme, with host masses of $\gtrsim 10^{13}\,M_\odot$ and a quasar lifetime approaching the Hubble time ($t_Q \approx 10^8 - 10^9$ yr) (Shen et al. 2007; Pizzati et al. 2024a). As highlighted by several studies (White et al. 2008; Wyithe & Loeb 2009; Shankar et al. 2010b), these values imply a steep and tight relation between the luminosity of quasars and the mass of the host halos, with SMBHs being either over-massive compared to their host halos/galaxies or having a large Eddington ratio (Pizzati et al. 2024a). While these trends need to be backed up by the higher signal-to-noise measurements that will be allowed by future optical large-sky surveys, they paint a very interesting picture and call for studies of quasar clustering at even higher redshifts.

Measurements of the quasar auto-correlation function at $z \gtrsim 5$, however, are extremely challenging due to the rapid decline of the quasar abundance at high redshift (e.g., Schindler et al. 2023). One alternative pathway to determine the clustering of quasars is to cross-correlate them with some other tracer, e.g., coeval galaxies. The idea behind these measurements is that, if we assume that both quasars and galaxies trace the same underlying dark matter density distribution, but with different bias factors, the cross-correlation function between these two classes of objects is entirely determined by their respective auto-correlation functions. Given that the clustering of high-z galaxies can be determined more easily due to their larger abundance, one can then measure the cross-correlation between quasars and galaxies (or, equivalently, study the over-densities of galaxies around quasars) to infer how strongly quasars are clustered in the high-z Universe.

Studies of the quasar-galaxy cross-correlation function are numerous at $z \approx 0 - 5$, with results that overall confirm an increase in the clustering strength with redshift (e.g., Adelberger & Steidel 2005; Shen et al. 2013; Ikeda et al. 2015; García-Vergara et al. 2017; He et al. 2018; García-Vergara et al. 2019). Nonetheless, two decades of ground- and space-based searches for galaxy over-densities around $z \gtrsim 6$ quasars have yielded mixed results, and contradictory claims have been made about the density (and clustering strength) of the primordial environments where these quasars live (e.g. Stiavelli et al. 2005; Willott et al. 2005; Zheng et al. 2006; Kim et al. 2009; Morselli et al. 2014; Simpson et al. 2014; Mazzucchelli et al. 2017a; Mignoli et al. 2020). In summary, even though the first studies on quasar clustering date back to more than two decades ago, extending these studies into the first billion years of cosmic history – where the link between quasar clustering and SMBH growth is even more relevant – has been extremely challenging.

Recently, however, ground-breaking progress has been made following both of the two independent pathways mentioned above. Exploiting the high sensitivity of the Subaru High-z Exploration of Low-Luminosity Quasars (SHELLQs) survey, Arita et al. (2023) have compiled a sample of ≈ 100 faint quasars at $z \approx 6$ and measured for the first time the large-scale quasar auto-

correlation function at those redshifts. Despite the large uncertainties due to the limited size of their sample, the authors measured an auto-correlation length of $r_{0,\mathrm{QQ}} = 24 \pm 11\,\mathrm{cMpc}\,h^{-1}$, in line with the trend observed at $z \approx 4$.

The launch of JWST, on the other hand, has opened up the possibility of obtaining large statistical samples of spectroscopically confirmed high-redshift galaxies, thus promising to revolutionize the search for over-densities around $z \approx 6$ quasars. Indeed, several independent studies (Kashino et al. 2023; Wang et al. 2023) have already used NIRCam Wide Field Slitless Spectroscopic (WFSS) observations of $z \approx 6$ quasar fields to show that these quasars reside in cMpc-scale over-densities traced by [O III] -emitting galaxies ([O III] *emitters*). Leveraging these unprecedented capabilities of JWST in studying the clustering and large-scale environment of high-redshift quasars, Eilers et al. (2024, hereafter E24) used observations from the EIGER survey (Kashino et al. 2023; Matthee et al. 2023) to compile a catalog of [O III] emitters in the environments of four bright $z \approx 6$ quasars, and measured for the first time the quasar-galaxy cross-correlation function at the same redshift. By also measuring the galaxy auto-correlation function, the authors concluded that high-z quasars live on average in $\approx 10^{12.3}\,\mathrm{M}_\odot$ halos, although with a substantial quasar-to-quasar variance in terms of environments. This finding implies that $z \approx 6$ quasars typically reside in moderately strong over-densities but not necessarily in the rarest and most massive environments that are present in the early Universe.

These measurements of the $z \approx 6$ quasar auto-/cross-correlation functions offer a unique opportunity to study SMBHs and their properties at high-z. In Pizzati et al. (2024a) (Chapter 2; hereafter, P24), we showed that quasar clustering measurements can be combined with quasar demographic properties (expressed by the quasar luminosity function, QLF) to infer fundamental quantities such as the quasar luminosity-halo mass relation, the mass function of halos that host active quasars (the quasar-host mass function, QHMF), the quasar duty cycle and the quasar lifetime. P24 makes use of a novel method that combines the outputs of dark-matter-only (DMO) cosmological simulations (specifically, the halo mass function and the cross-correlation function of halos with different masses) with an empirical quasar population model founded on a conditional luminosity function (CLF) framework (e.g., Yang et al. 2003). The authors applied this model to measurements of the quasar auto-correlation and quasar luminosity functions at $z \approx 2-4$, tracing the rapid change in SMBHs properties taking place at those redshifts.

In this work, we aim to extend the P24 model to interpret the new measurements of the quasar-galaxy cross-correlation function and the auto-correlation functions of quasars and galaxies at $z \approx 6$. These clustering measurements encompass a wide range of scales ($10^{-1} \lesssim r/\mathrm{cMpc} \lesssim 10^3$) and quasar luminosities ($10^{45.5} \lesssim L/\mathrm{erg\,s^{-1}} \lesssim 10^{48}$). Even more relevantly, modeling $z \approx 6$ galaxies and quasars simultaneously to compute their cross-

correlation statistics means that we must describe objects whose abundances span more than seven orders of magnitude (Schindler et al. 2023; Matthee et al. 2023). To overcome these obstacles, we extended the FLAMINGO suite (Schaye et al. 2023; Kugel et al. 2023) with a new 2.8 cGpc dark-matter-only simulation evolving more than a trillion particles and reaching the same resolution as the previous FLAMINGO DMO high-resolution runs (Schaye et al. 2023) but in a much larger volume. By employing this new, state-of-the-art, N-body simulation, named FLAMINGO-10k, we have the capability of modeling the clustering and demographic properties of quasars and galaxies simultaneously, providing a simple but powerful framework to interpret the large-scale environments of quasars and the properties of SMBHs in the first billion years of cosmic history.

The paper is structured as follows. In Sec. 3.2, we summarize the main features of the P24 model and describe the improvements performed in this work. Sec. 3.2.1 lays down the general theoretical framework while the new FLAMINGO-10k simulation is described in Sec. 3.2.2. Sec. 3.3 describes the comparison of our model with observational data, and Sec. 3.4 presents the main results of our analysis. These results are discussed and interpreted in the framework of current SMBH formation and evolution theories in Sec. 3.5. Conclusions are provided in Sec. 3.6.

3.2 Methods

The P24 model takes two fundamental ingredients from cosmological simulations, i.e. the halo mass function and the cross-correlation functions of halos with different masses, and combines these with a quasar conditional luminosity function (which stochastically assigns quasars to halos) to reproduce observations of the quasar luminosity function and the quasar auto-correlation function, together with other relevant quantities such as the mass function of quasar-hosting halos and the quasar duty cycle (see Fig. 2.1 for an overview).

Here, we plan to adapt this framework to include the presence of galaxies in the model, with the aim of reproducing their clustering and demographic properties in conjunction with the ones of quasars. We introduce the quasar-galaxy population modeling in Sec. 3.2.1 and Appendix 3.A, and present the FLAMINGO-10k simulation on which the model is founded in Sec. 3.2.2.

3.2.1 Quasar and galaxy population models

The primary goal of our model is to reproduce observations of the luminosity function and the clustering for both galaxies and quasars. In Appendix 3.A, we outline a general framework that allows us to use a conditional luminosity function (CLF) to stochastically connect dark matter halos to

any population of objects that are tracers of the underlying halo distribution and emit radiation with some luminosity, L. As discussed in the Appendix, both quasars and galaxies are suitable tracers to which this framework can be applied. We do so simultaneously: we define a conditional luminosity function for quasars, $\mathrm{CLF}_{\mathrm{QSO}}(L|M)$, and one for galaxies, $\mathrm{CLF}_{\mathrm{Gal}}(L|M)$ – with L being the luminosity of quasars/galaxies and M the mass of the host halos.

It is important to note that our definition of quasars and galaxies is entirely empirical, and it is solely based on our objective to reproduce a specific set of observations concerning these sources (see Introduction and Sec. 3.3.1). For this reason, our quasar population model is intended to describe only UV-bright, type-I quasars (e.g., Padovani et al. 2017). As for galaxies, our objective is to match JWST observations of [O III] emitters (E24), and thus – when not explicitly stated otherwise – we will use the words "galaxies" to describe only the ones that are bright in [O III] . Nonetheless, we stress the fact that the framework presented here is general and can be extended to different sub-populations of quasars/galaxies.

Another important note concerns the luminosity, L, of quasars and galaxies, which can also be set to any arbitrary choice (e.g., the bolometric luminosity or the luminosity of a specific line/band). As also done in P24, we choose to work with bolometric luminosities when modeling quasars. Therefore, the quasar conditional luminosity function, $\mathrm{CLF}_{\mathrm{QSO}}(L|M)$, will link the mass of host halos to the bolometric luminosities of quasars (i.e., $L \equiv L_{\mathrm{bol}}$). For galaxies, we use the luminosity of the $[\mathrm{O\ III}]_{5008}$ line, L_{OIII} instead, as this is the quantity that determines the detectability of the galaxies in the (slitless) JWST surveys. Therefore, $\mathrm{CLF}_{\mathrm{Gal}}(L|M)$ relates halos to [O III] luminosities (i.e., $L \equiv L_{\mathrm{OIII}}$). In the following, we will always use the symbol L, but add the caveat that the specific value of this symbol is different depending on whether we refer to quasars or galaxies.

We assume the same functional form for the two conditional luminosity functions, $\mathrm{CLF}_{\mathrm{QSO}}$ and $\mathrm{CLF}_{\mathrm{Gal}}$. Following P24, we write[1]

$$\mathrm{CLF}_i(L|M)\,\mathrm{d}L = \frac{f_{\mathrm{on}}^{(i)}}{\sqrt{2\pi}\sigma^{(i)}}\, e^{-\frac{\left(\log_{10} L - \log_{10} L_{\mathrm{c}}^{(i)}(M)\right)^2}{2\sigma^{(i)2}}} \mathrm{d}\log_{10} L, \tag{3.1}$$

where i stands either for "QSO" or "Gal". The characteristic luminosity, $L_{\mathrm{c}}^{(i)}$, has a power-law dependence on halo mass:

$$L_{\mathrm{c}}^{(i)}(M) = L_{\mathrm{ref}}^{(i)} \left(\frac{M}{M_{\mathrm{ref}}}\right)^{\gamma^{(i)}}. \tag{3.2}$$

[1] As also discussed in P24, the factor f_{on} accounts for the fact that not all quasars/galaxies may be luminous at any given time. In other words, we are implicitly assuming that a fraction of sources are inactive or simply too dim to be revealed by any observations and therefore we do not include their contribution in the CLF.

with $M_{\rm ref}$ being a reference mass that is associated with the reference luminosity $L_{\rm ref}$; we fix it to $\log_{10} M_{\rm ref}/{\rm M}_\odot = 12.5$. The free parameters of the model, which we will infer directly from observations in Sec. 3.3.1-3.4, are $\sigma^{\rm (QSO,Gal)}$, $L_{\rm ref}^{\rm (QSO,Gal)}$, $\gamma^{\rm (QSO,Gal)}$, and $f_{\rm on}^{\rm (QSO,Gal)}$. Note that, as in P24, we assume that these parameters do not depend on other variables such as halo mass or quasar luminosity.

Using the general framework outlined in Appendix 3.A (see also P24), we can combine each conditional luminosity function, ${\rm CLF_{QSO}}$ and ${\rm CLF_{Gal}}$, with the halo mass function, $n_{\rm HMF}$, to obtain fundamental quantities describing quasars and galaxies, such as their luminosity functions ($n_{\rm QLF}$ and $n_{\rm GLF}$), host mass functions ($n_{\rm QHMF}$ and $n_{\rm GHMF}$), and duty cycles ($\varepsilon_{\rm QDC}$ and $\varepsilon_{\rm GDC}$).

The quasar luminosity function (QLF) and the galaxy luminosity function (GLF) are observable quantities, and hence the predictions from our model for these functions can be directly compared with data. As for the quasar-host mass function (QHMF) and the galaxy-host mass function (GHMF), they determine the clustering properties of quasars and galaxies, respectively.

In particular, we follow here the approach described in P24 (see their Section 1 and Appendix A) to write the clustering properties of a population of objects given its host halo mass distribution. This approach assumes that the cross-correlation functions of dark matter halos with different masses are known. We describe in Sec. 3.2.2 and Appendix 3.B how to extract these cross-correlation terms from a cosmological simulation. Here, we assume that, after creating bins in halo mass, we can write the cross-correlation between two mass bins as $\xi_h(M_j, M_k; r)$, with $M_{j,k}$ being the bin centers.

The point made in P24 is that all the correlation functions concerning quasars and galaxies are simply weighted averages of these cross-correlation terms, with the weights (Q_j, G_j) determined by the specific host mass distribution we are considering ($n_{\rm QHMF}$ for quasars and $n_{\rm GHMF}$ for galaxies). In particular, we can define the weights Q_j to be:

$$Q_j = \frac{n_{\rm QHMF}(M_j|L > L_{\rm thr})\,\Delta M}{\int_0^{M_{\rm max}} n_{\rm QHMF}(M|L > L_{\rm thr})\,{\rm d}M}, \tag{3.3}$$

with ΔM being the width of the mass bins. The identical weighting for galaxies, G_j, reads:

$$G_j = \frac{n_{\rm GHMF}(M_j|L > L_{\rm thr})\,\Delta M}{\int_0^{M_{\rm max}} n_{\rm GHMF}(M|L > L_{\rm thr})\,{\rm d}M}. \tag{3.4}$$

With these definitions, we can write all correlation functions in the general form (with A and B representing two different populations of halo tracers):

$$\xi_{\rm AB}(r) = \sum_{j,k} A_j B_k \xi_h(M_j, M_k; r). \tag{3.5}$$

This expression implies that the quasar auto-correlation function, $\xi_{\rm QQ}(r)$, can simply be written as:

$$\xi_{\rm QQ}(r) = \sum_{j,k} Q_j Q_k \xi_h(M_j, M_k; r), \tag{3.6}$$

with the weights set by eq. 3.3. In the same way, the galaxy auto-correlation function, $\xi_{\rm GG}(r)$, reads:

$$\xi_{\rm GG}(r) = \sum_{j,k} G_j G_k \xi_h(M_j, M_k; r), \tag{3.7}$$

Finally, the cross-correlation function between quasars and galaxies, $\xi_{\rm QG}(r)$, is retained by weighting over the QHMF and the GHMF simultaneously:

$$\xi_{\rm QG}(r) = \sum_{j,k} Q_j G_k \xi_h(M_j, M_k; r). \tag{3.8}$$

As a final step, all of these correlation functions can be integrated along the line of sight direction to average out the contribution of redshift space distortions. In this way, we compute quantities that can be directly matched with data, such as the projected correlation function, $w_p(r_p)$, or the volume-averaged correlation function, $\chi_V(r_p)$. The former follows from a simple integration along the line of sight direction, π, with a limit $\pi_{\rm max}$ that is chosen according to observations:

$$w_p(r_p) = 2 \int_0^{\pi_{\rm max}} \xi(r_p, \pi)\, {\rm d}\pi, \tag{3.9}$$

while the latter implies that we choose a radial binning in the perpendicular direction, r_p, and a maximum distance in the parallel direction, $\pi_{\rm max}$, and perform a spatial average of the correlation function on every cylindrical bin. If we define $r_{p,\rm min}$ and $r_{p,\rm max}$ as the lower and upper limits of the radial bins, respectively, $\chi_V(r_p)$ can be simply expressed as:

$$\chi_V(r_p) = \frac{2}{V} \int_{r_{p,\rm min}}^{r_{p,\rm max}} \int_0^{\pi_{\rm max}} \xi(r_p, \pi)\, 2\pi r_p\, {\rm d}r_p\, {\rm d}\pi. \tag{3.10}$$

3.2.2 Simulation setup

As described in P24 (see Figure 2.1), we use dark-matter-only (DMO) cosmological simulations to extract two fundamental quantities that are at the core of our model: the *halo mass function*, $n_{\rm HMF}$, and the *cross-correlation functions* of halos with masses M_j and M_k, $\xi_h(M_j, M_k; r)$.

P24 used multiple simulations with different box sizes and resolutions to extend the range of masses that can be reliably modeled in their framework.

The argument in support of this approach was that every different simulation can describe the demographic and clustering properties of halos in a different range of masses, and putting together these properties allows for an exploration of a larger set of quasar-host mass distributions. This approach was particularly suited for getting an estimate of the quasar auto-correlation function, as this quantity primarily depends on the auto-correlation function of the halos whose mass is the maximum of the QHMF. For this reason, resolving very low and very high mass halos in the same simulation was not necessary, and the terms of the cross-correlation functions $\xi_h(M_i, M_j; r)$ with, e.g., $M_i \gg M_j$ were just extrapolated by appropriate analytic functions (see P24 for more details).

The problem we are facing here, however, is intrinsically different, as we need to model the cross-correlation function between quasars – which are very rare and are expected to live in massive halos – and galaxies – which are much more abundant and hence are hosted by much more common systems. This implies that the cross-correlation functions between very massive and less massive halos are at the core of our model, and hence they need to be faithfully represented in our numerical setup. For this reason, we use here a single simulation with a larger number of particles, intending to represent in the same box halos whose range of masses is broad enough to account for the presence of quasars and galaxies simultaneously. In the following, we give more details about the properties of this simulation, and we then proceed to describe how we extract from the simulated box the halo properties that our population models require.

3.2.2.1 Extending the suite of FLAMINGO runs: FLAMINGO-10k

FLAMINGO (Schaye et al. 2023; Kugel et al. 2023) is a suite of state-of-the-art, large-scale structure cosmological simulations combining hydrodynamical and dark-matter-only (DMO) runs in large volumes (≥ 1 Gpc). The simulations were performed using the coupled Particle-Mesh & Fast-Multipole-Method code SWIFT (Schaller et al. 2024). The fiducial runs adopt the "3x2pt + all" cosmology from Abbott et al. (2022) ($\Omega_\mathrm{m} = 0.306$, $\Omega_\mathrm{b} = 0.0486$, $\sigma_8 = 0.807$, $\mathrm{H}_0 = 68.1\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, $n_\mathrm{s} = 0.967$), with a summed neutrino mass of 0.06 eV. Initial conditions (ICs) are set using multi-fluid third-order Lagrangian perturbation theory (3LPT) implemented in MONOFONIC (Hahn et al. 2020; Michaux et al. 2021). Partially fixed ICs are used to limit the impact of cosmic variance (Angulo & Pontzen 2016) by setting the amplitudes of modes with wavelengths larger than 1/32 of the simulation volume side-length to the mean. The most demanding simulation in the suite (the `L2p8_m9` run of Schaye et al. 2023) encompassed a volume of side-length 2.8 cGpc with particles of mass $6.72 \times 10^9\,\mathrm{M}_\odot$.

Whilst the volume of this flagship run is sufficient for the present study, the resolution is not high enough to reliably characterize the halo mass and clustering of the [OIII] emitters we seek to study. We thus ran an additional simulation, FLAMINGO-10k, which we add to the FLAMINGO suite. FLAMINGO-10k was run on 65 536 compute cores, using the same setup (software, cosmology, ...) as the previous DMO FLAMINGO simulations, but with 8x higher resolution than the `L2p8_m9` run and a higher starting redshift ($z = 63$). The box size of this new simulation is chosen according to the flagship FLAMINGO run, $L = 2.8\,\mathrm{cGpc}$, while the resolution of the simulation reaches the one of the 1 cGpc FLAMINGO DMO high-resolution run ($m_{\mathrm{CDM}} = 8.40 \times 10^8\,\mathrm{M}_\odot$). The simulation makes use of 10080^3 cold dark matter (CDM) particles and 5600^3 neutrino particles, resulting in a total number of particles close to 1.2×10^{12}. As detailed in Sec. 3.4, this large number of particles will let us model halos whose masses span more than two orders of magnitude at $z \approx 6$ throughout the $(2.8\,\mathrm{cGpc})^3$ volume. The particles and halo catalogs were stored at 145 redshifts between $z = 30$ and $z = 0$ with 31 outputs at $z > 6$, allowing for the precise tracing of the growth of structures at early times.

3.2.2.2 Obtaining the sub-halo catalogue with HBT+

The first step that we take once we have the final simulated volume is to build a halo catalogue containing the positions and masses of all (sub-)halos in the simulation. In P24, we included only central halos in the catalogue and discarded the contribution of satellite haloes completely. This was done because our main focus was the auto-correlation function of quasars at large scales ($r \gtrsim 5\,\mathrm{cMpc}$). Here, instead, we aim to reproduce correlation functions down to $r \approx 0.1\,\mathrm{cMpc}$ (i.e., well within the virial radii of massive halos), and hence the contribution of all sub-haloes must be carefully considered. We note that in our framework (Sec. 3.2.1) we do not make any explicit distinctions between central sub-halos and satellites. For this reason, we build a halo catalogue that includes all kinds of sub-halos, and we use the general term "halo" to refer to any kind of sub-halos, irrespective of whether they are central or satellite. In general, whenever we refer to quasar/galaxy hosts in the context of our model (e.g., in the QHMF and GHMF), we always implicitly assume that we are talking about *sub-halos*, and not about the larger groups identified by a friends-of-friends algorithms.

We select a single snapshot from FLAMINGO-10k at $z = 6.14$, which represents the closest match in terms of redshift to the observations we aim to reproduce in this work (Sec. 3.3.1). We use this snapshot together with all the other ones at higher-z to build a halo catalogue using the upgraded Hierachical Bound-Tracing (HBT+) code (Han et al. 2012, 2018). HBT+ identifies sub-haloes as they form and tracks their evolution as they merge. By consistently following sub-haloes across cosmic times HBT+ represents a

robust solution to the problem of identifying small-scale bound structures in DMO simulations. This is the ideal choice for the problem we are facing here, as we aim to represent the spatial distribution of quasars and galaxies down to very small spatial scales.

We use the bound mass definition for (sub-)halo masses. In other words, we compute the mass of each (sub-)halo by summing up the mass of all its bound particles. Since tidal stripping decreases the mass of satellite halos by a significant amount, we use here the peak halo mass, $M_{\rm peak}$, which is defined as the largest bound mass that a (sub-)halo has had across cosmic history. In practice, HBT+ saves this mass for each snapshot, and so we can simply use the peak bound masses that are given in the output by the code for our population model (i.e., $M \equiv M_{\rm peak}$). We then complete the catalogue by adding the position of each (sub-)halo, which we define by looking at its centre of potential.

3.2.2.3 A simulation-based analytical description of halo properties

Once we have obtained a catalogue with the positions and masses of halos in the simulation at a given redshift, we can easily compute the halo mass function and the (cross-)correlation functions of halos with different masses. However, as also done in P24, we aim to describe these quantities with analytical functions, which we fit to the outputs of the simulation. This approach allows us to obtain a very general description of halo properties, independent of the specific mass bins employed. More importantly, in P24 we have shown that using these fitting functions we can smoothly extrapolate the behavior of the cross-correlation functions even to the combinations of mass bins for which there are very few halos available in the simulation, and hence for which the correlation functions measured numerically are extremely noisy and uncertain. This simple step improves the quality of our parameter inference (Sec. 3.3) and lets us recover well-behaved posterior distributions for a wide range of model parameters.

Fitting the halo mass function is straightforward. As in P24, we consider the same functional form used by Tinker et al. (2008) (see also Jenkins et al. 2001; White 2001; Warren et al. 2006) for the fit, and consider all halos above the minimum mass $\log M_{\rm min}/{\rm M}_\odot = 10.5$, corresponding to halos with more than ≈ 40 particles.

As for the cross-correlation function of halos with mass M_j and M_k, $\xi_h(M_j, M_k; r)$, we first compute each correlation function numerically by creating a grid in mass and distance made by 8 uniformly spaced bins in $\log_{10} M$, with a minimum halo mass of $\log_{10} M_{\rm min}/{\rm M}_\odot = 10.5$ and a maximum of $\log_{10} M_{\rm max}/{\rm M}_\odot = 12.5$, and 18 (logarithmically-spaced) bins in the radial direction with a minimum radial distance of $\log_{10} r_{\rm min}/{\rm cMpc} = -1$ and a maximum of $\log_{10} r_{\rm max}/{\rm cMpc} = 2.2$. We then use the package

CORRFUNC (Sinha & Garrison 2020) to compute the number of halo pairs in the simulated catalogues for every combination of masses and distance. We use a simple estimator to obtain the halo cross-correlation functions:

$$\xi_h(M_j, M_k; r) = \xi_{j,k}(r) = \frac{D_j D_k(r)}{R_j R_k(r)} - 1, \tag{3.11}$$

where $D_j D_k$ stands for the number of pairs of halos in the mass bin j with halos in the mass bin k, whereas $R_j R_k$ refers to the number of pairs when comparing to a random distribution of the same halos. For a periodic box of volume V, $R_j R_k$ can be simply expressed analytically as:

$$R_j R_k = \frac{4\pi}{3V} \left(r_{\rm max}^3 - r_{\rm min}^3\right) N_j N_k, \tag{3.12}$$

with N_j and N_k being the number of halos in the mass bins j and k, respectively, and $r_{\rm min,max}$ the limits of the radial bin considered.

We fit $\xi_h(M_j, M_k; r)$ with the same setup as described in P24. In short, we divide all the cross-correlation terms, $\xi_h(M_j, M_k; r)$, by a reference correlation function, $\xi_{\rm ref}(r)$, which we set equal to the auto-correlation function of the first mass bin. Then, we fit the resulting functions with a 3-d polynomial to capture the residual dependencies on the two masses and the distance. The fit is performed by converting masses to peak heights, $\nu(M) = \delta_c/\sigma(M, z)$ – with $\delta_c \approx 1.69$ and $\sigma^2(M, z)$ being the variance of the smoothed linear density field (see also Sec. 3.5.1). We adopt this approach in order to minimize any dependences of the cross-correlation functions on cosmology and redshift. Errors on the cross-correlation terms are chosen by assuming Poissonian uncertainties on the halo pair counts. Finally, we note that, before fitting, we weigh every uniform mass bin with the halo mass function, so that the effective mass M_k corresponding to the bin k is not the bin center, but the median value of the halo mass function in that specific bin.

After performing the fit, we introduce here a further step that aims to achieve a better description of the cross-correlation functions at large scales, $r \gtrsim 20 - 40\,$cMpc. As noted in P24, the values of the correlation functions extracted from simulations tend to be unreliable at large scales for two reasons. First, the finite size of the box reduces the number of very large-scale pairs that are available. Secondly, at $r \gtrsim 100\,$cMpc the behavior of correlation functions becomes non-trivial due to the presence of the baryon acoustic oscillations (BAO) peak, which is hard to capture with the coarse binning employed here. At large scales, however, density perturbations are linear and they can faithfully be described by the linear halo bias framework (Bardeen et al. 1986; Cole & Kaiser 1989; Jing 1998; Cooray & Sheth 2002). For this reason, we follow Nishimichi et al. (2021) and smoothly interpolate between our fit to simulations at small-to-medium

scales and the predictions from linear theory at large scales. In practice, we introduce a damping function $D(r)$, and write the correlation functions $\xi_h(M_j, M_k; r)$ as:

$$\xi_h(M_j, M_k; r) = D(r)\xi_{h,\mathrm{fit}}(M_j, M_k; r) + (1 - D(r))\xi_{h,\mathrm{lin}}(M_j, M_k; r), \quad (3.13)$$

where $\xi_{h,\mathrm{fit}}(M_j, M_k; r)$ is the fit performed to simulations described above, while $\xi_{h,\mathrm{lin}}(M_j, M_k; r)$ is the prediction coming from the linear halo bias framework (based on linear theory, see, e.g., Murray et al. 2021):

$$\xi_{h,\mathrm{lin}}(M_j, M_k; r) = b(M_j)b(M_k)\xi_{\mathrm{mm}}(r). \quad (3.14)$$

We use the package COLOSSUS (Diemer 2018) to compute the matter auto-correlation function, $\xi_{\mathrm{mm}}(r)$, and the linear bias factors, $b(M_{j,k})$, based on the Tinker et al. (2010) relation. As for the damping function, we choose the following functional form:

$$D(r) = e^{-\left(\frac{r}{r_{\mathrm{lin}}}\right)^{\alpha}}, \quad (3.15)$$

with the parameters set to $\alpha = 5$ and $r_{\mathrm{lin}} = 20\,\mathrm{cMpc}$.

In summary, we adopt here an extension of the P24 fitting framework that uses DMO simulations to provide an analytical description of the demographic and clustering properties of halos, expressed by the halo mass function and the halo cross-correlation functions. Thanks to the use of fitting functions, we can extrapolate the behavior of these quantities for a very large range of masses (from $\log M/\mathrm{M}_\odot \approx 10.5$ to $\log M/\mathrm{M}_\odot \approx 13 - 13.5$), and, by smoothly interpolating between DMO simulations at small scales and linear theory at large scales, our correlation functions can capture more than four orders of magnitude in scale (from $r \approx 0.1\,\mathrm{cMpc}$ out to $r \approx 1\,\mathrm{cGpc}$). As shown in the following Section, these properties are essential to reproduce the large diversity of data concerning galaxies and quasars that are the focus of the present work.

In Appendix 3.B, we show the results for the fit of the cross-correlation function terms and elaborate on the validity of this approach in the context of our analysis. Further discussion on the general methodology employed here can be found in P24.

3.3 Data-Model comparison

Adopting the methodology described in the previous Section, we can obtain all the ingredients needed to compare our model with observational data. The model depends on eight free parameters (see Sec. 3.2.1), that we constrain by jointly fitting the luminosity and clustering measurements of both quasars and galaxies. We provide a brief description of the data considered in the analysis in Sec. 3.3.1, and proceed to the comparison with our model in Sec. 3.3.2.

Table 3.1: Summary of all the data we compare our model with, together with a quantitative measurement (χ^2 statistics) of the quality of the fit. The χ^2 is computed by considering the best-fit parameters coming from the joint fit (see main text), and n stands for the number of data points that we fit for each quantity. A discussion on the quality of the fit can be found in Sec. 3.3.2. References are: (a) Schindler et al. (2023); (b) Matthee et al. (2023); (c) E24; (d) Arita et al. (2023).

Name	Quantity	Survey Name	Redshift Range	Fig.	Ref.	χ^2/n
Quasar luminosity function	$n_{\rm QLF}(L)$	PS1, SHELLQs	$5.7-6.2$	3.2 (top)	(a)	8.3/10
Galaxy luminosity function[a]	$n_{\rm GLF}(L)$	EIGER	$5.3-6.9$	3.2 (top)	(b)	8.7/7
Quasar-Galaxy cross-corr. function	$\chi_{V,\rm QG}(r_p)$	EIGER	$5.9-6.4$	3.2 (bottom)	(c)	7.4/8
Galaxy auto-corr. function	$\chi_{V,\rm GG}(r_p)$	EIGER	$5.3-6.9$	3.2 (bottom)	(c)	15.1/8
Quasar auto-corr. function	$w_{p,\rm QQ}(r_p)/r_p$	SHELLQs	$5.8-6.6$	3.8	(d)	6.1/5[b]

a We exclude the innermost bin because it is very uncertain due to low completeness.
b This dataset is excluded from the joint fit, and analysed separately in Appendix 3.C. The χ^2 reported here is the value obtained using the best-fit parameters coming from the joint fit of all the other datasets.

3.3.1 Overview of observational data

The data we consider in this work concern the luminosity functions and auto-correlation functions of quasars and galaxies, and the cross-correlation function between these two different populations. In Table 3.1, we summarize all these data and point to their respective references. The $z \approx 6$ quasar luminosity function (QLF) is taken from Schindler et al. (2023), and it is compiled including 125 quasars with $-28 \lesssim M_{1450} \lesssim -25$ from the the Pan-STARRS1 (PS1) quasar survey (Bañados et al. 2016), as well as 48 fainter ($-25 \lesssim M_{1450} \lesssim -22$) quasars from the SHELLQs survey (Kashikawa et al. 2015; Matsuoka et al. 2018). Note that, as detailed in Sec. 3.2.1 and in P24, we convert absolute magnitudes to quasar bolometric luminosities using the relation from Runnoe et al. (2012a)[2]. The galaxy luminosity function (GLF), based on JWST observations of [O III] emitters, was compiled by Matthee et al. (2023) in the context of the EIGER survey. The luminosities of galaxies are already expressed in [O III] line fluxes, in accordance with our population model (Sec. 3.2.1), and cover the range $42 \lesssim \log_{10} L_{\rm OIII}/{\rm erg\,s^{-1}} \lesssim 43.5$. We discard the faintest bin in the GLF because, as discussed in Matthee et al. (2023), its completeness is relatively low ($\approx 40\%$), and hence the value of the abundance of galaxies in that bin is particularly uncertain.

The quasar-galaxy cross-correlation function and the galaxy auto-correlation function are also measured by the EIGER survey in E24. They both span a spatial range $0.1 \lesssim r/{\rm cMpc} \lesssim 6$, sharing the same radial bins. Being obtained with the same methodology and in the same analysis, these two datasets are homogeneous, and it is natural to consider them jointly. The quasar auto-correlation function (Arita et al. 2023), on the other hand, comes from a very different dataset: it includes quasars with much fainter luminosities from the SHELLQs survey (Matsuoka et al. 2018), and it constrains their clustering only at very large scales ($r \gtrsim 40\,{\rm cMpc}$; see Arita et al. 2023). Further discussion on this can be found in Sec. 3.3.2 and in Appendix 3.C.

One of the key aspects to bear in mind when analysing data concerning correlation functions is that our model is quite sensitive to the value of the luminosity threshold, $L_{\rm thr}$, considered when measuring quasar/galaxy clustering (see eq. 3.19-3.20). While properly modeling the effects of observational incompleteness in the context of our framework is beyond the scope of this work, it is important to set these threshold values carefully to ensure that we get unbiased results. Let us start with the E24 observations. The EIGER survey targets only five very bright quasars and detects galaxies in their fields. This implies that the quasar population whose clustering is being probed by EIGER consists only of very bright ($M_{1450} \lesssim -27$) sources.

[2]The bolometric correction for $\lambda = 1450$ Å is $\log_{10} L_{\rm iso}/{\rm erg\,s^{-1}} = 4.745 + 0.910 \log_{10} \lambda L_\lambda/{\rm erg\,s^{-1}}$. $L_{\rm iso}$ refers to the bolometric luminosity computed under the assumption of isotropy, and it is related to the real bolometric luminosity L through the relation $L = 0.75\, L_{\rm iso}$.

For this reason, we set a value of the quasar luminosity threshold for modeling the quasar-galaxy cross-correlation function of $\log_{10} L_{\rm thr,QSO}/{\rm erg\,s^{-1}} = 47.1$, which is consistent with the luminosity of the faintest quasar probed by EIGER. However, we mention the caveat that setting a luminosity threshold would only be possible for a luminosity-limited sample. In reality, the EIGER survey targets only a few selected quasar fields and is not constructed to reproduce the actual luminosity distribution of bright quasars. While this may introduce a minor bias in our results, we neglect this effect here and consider the EIGER sample to be representative of the $z \approx 6$ bright ($L > L_{\rm thr,QSO}$) quasar population.

As for galaxies, the minimum [O III] luminosity that EIGER measurements consider is $\log_{10} L/{\rm erg\,s^{-1}} \approx 42$. However, the sample starts to be significantly incomplete already at higher luminosities. This represents an issue in our framework, as the luminosity-halo mass relations assumed in eq. 3.2 imply that clustering is luminosity-dependent. Including a large population of low-luminosity galaxies of which only a fraction was detected in the observations because of low completeness would then bias our results, since the luminosity distribution of the galaxies for which clustering was measured would not be the same as the one resulting from our modeling by simply setting the luminosity limit to be the lowest luminosity considered. We can alleviate this problem by setting an effective luminosity threshold that accounts for the fact that the sample is largely incomplete at lower luminosities. We choose the following effective threshold for galaxies: $\log_{10} L_{\rm thr,Gal}/{\rm erg\,s^{-1}} \approx 42.4$. This value corresponds to the luminosity at which the average completeness of the EIGER sample drops below $\approx 75\%$ (Matthee et al. 2023). We employ an analogous argument to set the luminosity threshold for the quasar auto-correlation function measurements of Arita et al. (2023). We find the magnitude at which the completeness of the SHELLQs survey drops below 75%, and convert this magnitude to a quasar bolometric luminosity obtaining $\log_{10} L_{\rm thr,QSO}/{\rm erg\,s^{-1}} = 45.3$[3].

3.3.2 Parameter inference

We employ a Bayesian framework and write down the posterior distribution for the model parameters. As described in Sec. 3.2.1, the model has eight free parameters, describing the conditional luminosity functions of quasars and galaxies simultaneously. We choose the same parametrization for $\rm CLF_{QSO}$ and $\rm CLF_{Gal}$. As a result, the same sets of parameters account for the two functions: these are the normalization and slope of the quasar/galaxy luminosity-halo mass relation ($L_{\rm ref}$ and γ, respectively), the logarithmic scatter around this relation (σ), and the fraction of quasars/galaxies that

[3]As detailed in Sec. 3.3.2 and Appendix 3.C, we find that the data for the quasar auto-correlation function are not able to constrain our model parameters. For this reason, in this specific case, the value for the luminosity threshold we choose here is irrelevant.

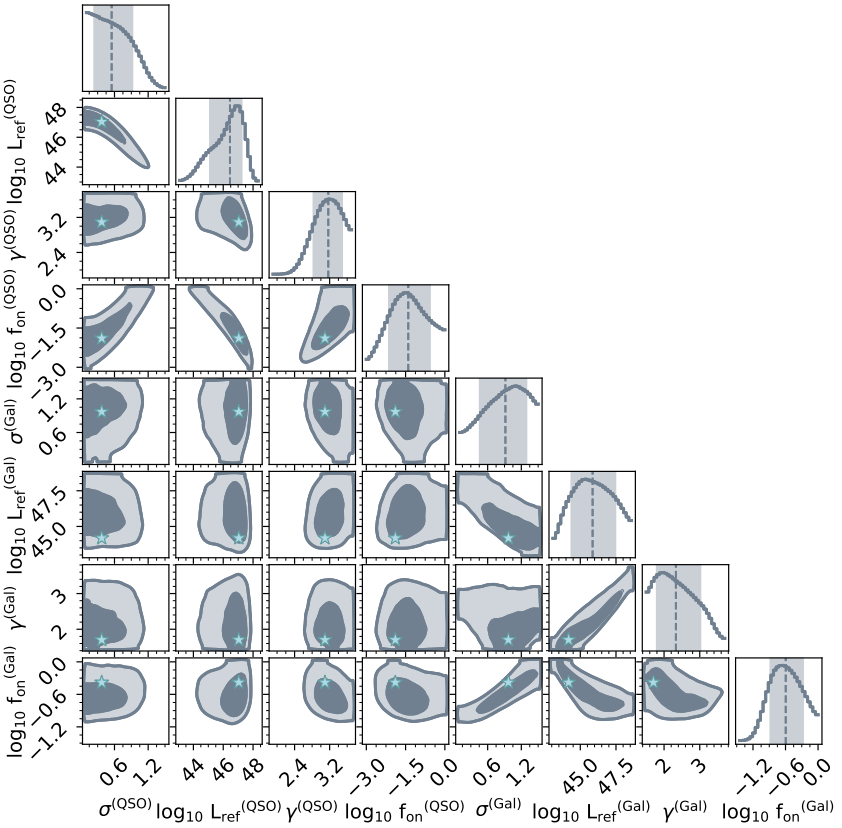

Figure 3.1: Corner plots of the 8-d posterior distribution for the joint fit described in Sec. 3.3.2. Contours in the 2-d histograms highlight the 1σ and 2σ regions, whereas the dashed lines in the 1-d histograms represent the median values of the parameters (with 1σ errors shown as shaded regions). The maximum-likelihood values are shown with star symbols in each corner plot. The units of the reference luminosity parameters $\log_{10} L_{\rm ref}^{\rm (QSO,Gal)}$ are $\mathrm{erg\,s^{-1}}$.

Table 3.2: Constraints (median values and 16th-84th percentiles) on the model parameters based on the corner plots shown in Figure 3.1. The eight parameters are divided between the ones describing the quasar CLF ("QSO") and the ones for the galaxy CLF ("Gal").

Quantity	σ	$\log_{10} L_{\rm ref}$ [$\mathrm{erg\,s^{-1}}$]	γ	$f_{\rm on}$ [%]
QSO	$0.55^{+0.37}_{-0.31}$	$46.45^{+0.79}_{-1.35}$	$3.17^{+0.32}_{-0.34}$	$3.9^{+21}_{-3.2}$
Gal.	$0.92^{+0.38}_{-0.46}$	$45.86^{+1.60}_{-1.49}$	$2.33^{+0.69}_{-0.54}$	25^{+31}_{-20}

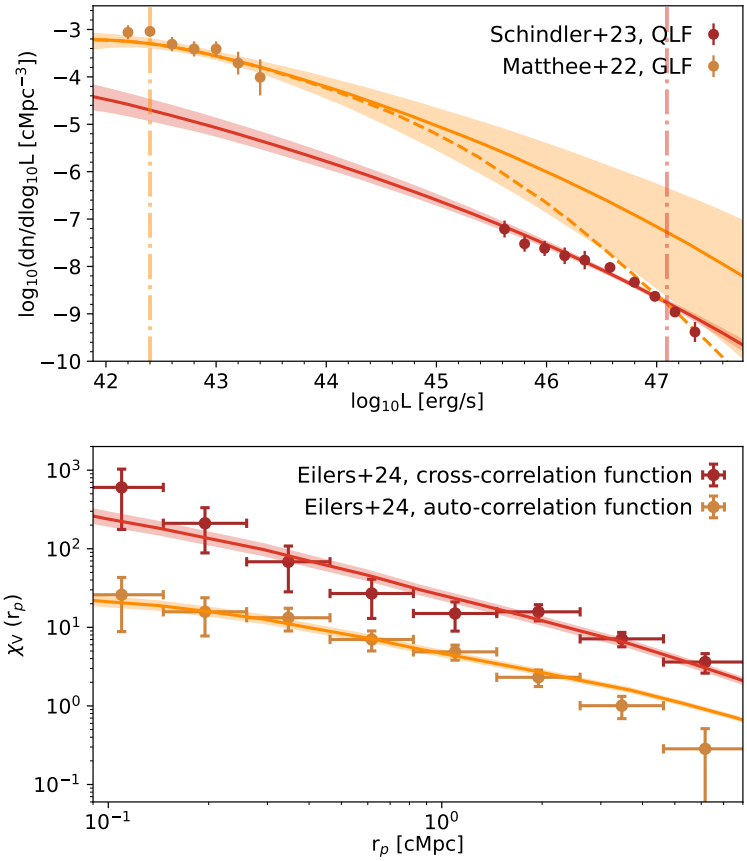

Figure 3.2: Comparison of the predicted luminosity (top) and correlation (bottom) functions with the observational data from Table 3.1. The galaxy luminosity function (GLF) and auto-correlation function are shown in orange, while the quasar luminosity function (QLF) and the quasar-galaxy cross-correlation function are shown in red. Median values (solid lines) and 1σ uncertainty regions (shaded areas) are obtained by randomly sampling the Markov chains for the posterior distribution 2000 times. The red and orange vertical dot-dashed lines in the upper right panel are the luminosity threshold for quasar and galaxies ($L_{\rm thr}$), respectively, that are used for modeling clustering measurements (see Sec. 3.3). The dashed line in the same panel represents the median value for the GLF when assuming that the galaxy luminosity-halo mass relation flattens for large halo masses (see Sec. 3.4 and Figure 3.3).

are active at any given moment ($f_{\rm on}$). The final set of parameters, Θ, is then: $(\sigma^{\rm (QSO)}, L_{\rm ref}^{\rm (QSO)}, \gamma^{\rm (QSO)}, f_{\rm on}^{\rm (QSO)}, \sigma^{\rm (Gal)}, L_{\rm ref}^{\rm (Gal)}, \gamma^{\rm (Gal)}, f_{\rm on}^{\rm (Gal)})$.

As in P24, we set flat uninformative priors on $\sigma^{\rm (QSO,Gal)}$ and $\gamma^{\rm (QSO,Gal)}$, and on the logarithm of $L_{\rm ref}^{\rm (QSO,Gal)}$ and $f_{\rm on}^{\rm (QSO,Gal)}$. We choose to explore a wide parameter space, letting the parameters vary with the following bounds: $\sigma^{\rm (QSO,Gal)} \in (0.1\,{\rm dex}, 2.0\,{\rm dex})$; $\log_{10} L_{\rm ref}^{\rm (QSO,Gal)}/{\rm erg\,s^{-1}} \in (43.0, 48.6)$; $\gamma^{\rm (QSO,Gal)} \in (1, 4)$; $\log_{10} f_{\rm on}^{\rm (QSO,Gal)} \in (-3, 0)$. The lower limits on $\sigma^{\rm (QSO,Gal)}$ and the upper limits $\log_{10} f_{\rm on}^{\rm (QSO,Gal)}$ are chosen because of physical constraints (i.e., the scatter in the $L - M$ relation is unlikely to be smaller than 0.1 dex and the active fraction is less than unity by definition).

We provide joint constraints on the parameters by fitting the data described in Sec. 3.3.1 (i.e., the luminosity and correlation functions for quasars and galaxies) simultaneously. In other words, we write the joint likelihood distribution as the product of the single Gaussian likelihoods for each dataset (we assume that all the measurements are independent):

$$\mathcal{L}^{\rm (joint)} = \prod_i \mathcal{L}^{\rm (i)}, \tag{3.16}$$

where i ranges over the datasets shown in Table 3.1.

When performing our analysis, we found that the data for the quasar auto-correlation function (Arita et al. 2023) were not able to place significant constraints on any of our model parameters. As a result, this dataset was not informative, and could not be used to infer any of the physical properties of quasars. This conclusion differs from the one found in Arita et al. (2023), where the authors are able to determine the range of host-halo masses for quasars at $z \approx 6$. We investigated the issue further and found that the different conclusions arise from different assumptions made for the shape of the auto-correlation functions at large scales. For this reason, we exclude the Arita et al. (2023) dataset from the joint fit performed here, and defer the analysis of this dataset to Appendix 3.C. In that Section, we compare in detail our analysis with the one performed by Arita et al. (2023) and conclude that, if we assume a physically-motivated choice for the shape of the quasar auto-correlation function, we are not able to place interesting constraints on the distribution of quasar-host halo masses.

Moving forward, we discuss the results of our parameter inference for the "joint" model described above, including all the other datasets compiled in Table 3.1. We explore the posterior distribution for this model using a Markov-Chain Monte Carlo (MCMC) approach. We employ the Python package EMCEE (Foreman-Mackey et al. 2013) to sample the posteriors using the affine-invariant ensemble prescription (Goodman & Weare 2010). We place $m = 48$ walkers distributed randomly in the parameter space and evolve them for $N = 10^5$ steps. Figure 3.1 shows the corner plot for the 8-d posterior distribution, while Table 3.2 summarizes the constraints we obtain

for each of the model parameters. The samples of the posterior distribution obtained by the Markov Chains are then used to obtain predictions for the luminosity and correlation functions, both for quasars and galaxies at the same time; we compare these quantities with the data in Figure 3.2. The top right panel shows predictions for the galaxy luminosity function (orange) and the quasar luminosity function (red), while the bottom panel shows the quasar-galaxy cross-correlation function (red) and the galaxy auto-correlation function (orange).

In all cases, we see that our model fares well when compared to the observational data. As a quantitative estimate of this accordance, we take the parameters corresponding to the maximum of the posterior distribution (highlighted by star symbols in Figure 3.1) and measure the χ^2 statistic for each of the single dataset shown in Table 3.1. Values of the χ^2 are reported in the last column of Table 3.1. We generally find a very good agreement between our model and every single dataset analyzed. The only exception is the galaxy auto-correlation function, for which the χ^2 is relatively large when compared to the size of the dataset. We believe this is due to the small reported uncertainties in the observational data, that are likely underestimated. As discussed in E24, these uncertainties are assigned according to the Poissonian statistics associated with the pair counts, and they do not take into account the uncertainty coming from cosmic variance as well as other possible systematic effects. This may be particularly relevant in the outermost bins, for which the data drop significantly more rapidly than what is predicted by our model. Covariance between different data points is also neglected in the E24 analysis, even though it most likely contributes to the total error budget significantly. This artificially increases the discrepancy between our model and the data.

3.4 Results

In the last Section, we have shown that we can successfully reproduce the data for the luminosity and correlation functions of quasars and galaxies with the simple extension of the P24 framework described in Sec. 3.2.1. In this framework, we use observations to constrain the conditional luminosity functions of quasars and galaxies simultaneously. In turn, each conditional luminosity function can be related to other fundamental properties such as the luminosity-halo mass relation, the host mass function, and the duty cycle/occupation fraction. We examine here these properties starting from the inferred values of the parameters obtained in Figure 3.1 and Table 3.2. We first examine quasar properties, and then turn our attention to galaxies.

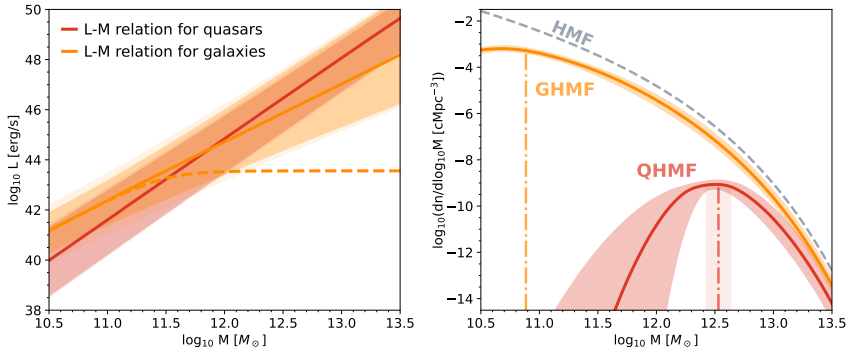

Figure 3.3: *Left*: Luminosity-halo mass relation for quasars (red) and galaxies (i.e., [O III] emitters; orange). The quasar luminosity is the bolometric one, while the galaxy luminosity is the one from the [O III] line. Median values (solid lines) and 16th-84th percentiles (dark-shaded areas) for the $L-M$ relations are obtained by randomly sampling the Markov chains for the posterior distribution 2000 times. The cumulative effects of the uncertainty on the median and the intrinsic scatter in the relations, expressed by the σ parameter in the CLF, are shown with a lighter shading. The dashed orange line corresponds to the modified galaxy luminosity-halo mass relation, with a flattening of the relation above a threshold mass of $M = 10^{11.5}\,\mathrm{M}_\odot$. *Right*: Quasar-host mass function (QHMF; red) and galaxy-host mass function (GHMF; orange). Median and 1σ uncertainties of these functions (obtained as in the left panel) are shown with solid lines and dark-shaded areas, respectively. The dashed-dotted lines show the median halo masses associated with the QHMF (red) and GHMF (orange) distributions (see eq. 3.21); light-shaded regions represent 1σ uncertainties on these median masses. The halo mass function (HMF) at the redshift of interest ($z = 6.14$) is shown with a gray dashed line.

3.4.1 The quasar luminosity-halo mass relation and the host halos of quasars at $z \approx 6$

Figure 3.3 shows the quasar luminosity-halo mass relation (left) and the quasar-host mass function (QHMF; right) at $z \approx 6$, as inferred from our model. We obtain a rather steep quasar $L - M$ relation, with a slope of $\gamma^{(\mathrm{QSO})} \approx 3.2$. This steep relation between quasar luminosities and halo masses is in agreement with the results of P24, which use data at $z \approx 2 - 4$ to study the evolution of this relation with redshift and find a significant increase in the slope parameter at earlier cosmic time. Our results suggest that this trend extends to even higher redshifts, with a close-to-linear relation at $z \approx 2$ turning into a very steep relation ($\gamma^{(\mathrm{QSO})} \approx 2 - 3$) at $z \approx 4 - 6$. We mention the caveat, however, that in this analysis the shape of the $L - M$ relation is primarily constrained by the QLF, and only marginally by the clustering measurements. This is because the E24 clustering data only focus on a very bright sub-sample of $z \approx 6$ quasars, and so they cannot constrain the behaviour of the $L - M$ relation below a luminosity of $\log_{10} L/\mathrm{erg\,s^{-1}} \approx 47$. Given that the shape and normalization of the QLF at high redshift are rather uncertain, especially at the faint end (e.g., Giallongo et al. 2019; Maiolino et al. 2024; Harikane et al. 2023; Andika et al. 2024), the shape of the $L - M$ relation is inevitably also plagued by this uncertainty.

The scatter in the quasar $L-M$ relation, on the other hand, is constrained both by the QLF and by the cross-correlation function simultaneously. In our analysis, we find a rather large log-normal scatter of $\sigma^{(\mathrm{QSO})} \approx 0.64$ dex (although with a significant uncertainty of ≈ 0.3 dex). This relatively large scatter is in line with the one measured by P24 at $z \approx 2.5$, but it represents a significant difference if compared to the very low scatter $\sigma^{(\mathrm{QSO})} \lesssim 0.3$ dex found by P24 at $z \approx 4$. Similarly, the value we obtain for the active fraction of $z \approx 6$ quasars $f_{\mathrm{on}}^{(\mathrm{QSO})}$ ($\approx 2\%$) is rather low if compared to the very high active fraction ($\approx 50\%$) found by P24 at $z \approx 4$. We defer the analysis of the peculiar redshift evolution traced by these parameters to Sec. 3.5.1.

The QHMF (Figure 3.3, right panel) reveals that quasars tend to live in $\log_{10} M/\mathrm{M}_\odot \approx 12.5$ halos (median value of $\log_{10} M/\mathrm{M}_\odot \approx 12.53 \pm 0.13$), with a rather broad distribution encompassing a large range of halo masses (from $\log_{10} M/\mathrm{M}_\odot \approx 12.1$ to $\log_{10} M/\mathrm{M}_\odot \approx 12.8$ at 2σ). The range of host masses we obtain is in perfect agreement with the conclusions of E24, who pointed out that quasars tend to live in moderately strong over-densities, but not necessarily in the most over-dense regions of the Universe (corresponding to halo masses of $\log_{10} M/\mathrm{M}_\odot \gtrsim 13$).

Even more interestingly, the broad distribution of host masses that we find from the inferred QHMF is compatible with the large quasar-to-quasar variance in terms of over-densities found by E24. The diversity of environments emerging from the E24 observations is likely a combination

of cosmic variance and variance in the host halo masses of quasars and/or galaxies. While we leave a quantitative analysis of these sources of variance to future work, it is encouraging to find evidence for the latter in our results. We stress the fact that our method for obtaining the QHMF does not make use of the observed diversity in terms of environments, as it only focuses on the global demographic and clustering properties of galaxies and quasars. The broad distribution of host masses that we find from our QHMF follows naturally from jointly modeling the clustering properties of quasars together with the shape and normalization of the quasar luminosity function.

In the analysis presented in E24, the framework developed here was used to match the quasar-galaxy cross-correlation function and the galaxy auto-correlation function by assuming simple "step-function" halo occupation distributions (HODs) for both quasars and galaxies. In other words, E24 populated halos and galaxies only above some minimum mass thresholds. With this method, they inferred the *minimum* host halo mass for quasars to be $\log_{10} M/\mathrm{M}_\odot \approx 12.43$. For a "step-function" HOD model, this value corresponds to a *median* quasar host mass of $\log_{10} M/\mathrm{M}_\odot \approx 12.51$, in excellent agreement with the median value of our inferred QHMF distribution.

Our conclusions on the quasar-host masses are also in line with the ones obtained by Mackenzie et al. in prep. In this work, the authors use the UniverseMachine (Behroozi et al. 2019) to compare the number of satellite halos to the number of companion galaxies observed in EIGER quasar fields. In this way, they obtain a distribution of possible host dark matter halo masses for each observed quasar in E24. Overall, the median value they obtain by putting together all the different mass distributions is $\log_{10} M/\mathrm{M}_\odot = 12.4 \pm 0.5$. The agreement with our results is significant, considering the very different assumptions underlying this method compared to the ones made here. Another estimate for the typical host halo masses of EIGER quasars was also obtained in E24 by comparing the observed $\chi_{V,\mathrm{QG}}$ with predictions from the TRINITY model (Zhang et al. 2023b). The resulting median host halo mass, $\log_{10} M/M_\odot = 12.14^{+0.24}_{-0.26}$, is slightly lower than the one found here, but still marginally compatible when taking uncertainties into account.

Finally, by relating the inferred QHMF to the halo mass function (HMF) at the same redshift (see eq. 3.22), we can obtain an estimate for the quasar duty cycle, $\varepsilon_{\mathrm{QDC}}$. Fig. 3.4 (left panel) shows the probability density function (PDF) for the quasar duty cycle (red) and the galaxy duty cycle (orange) obtained by randomly sampling the Markov chains for the posterior distribution shown in Fig. 3.1. We infer a value for the quasar duty cycle of $\varepsilon_{\mathrm{QDC}} = 0.9^{+2.3}_{-0.7}\%$. This relatively low value of the duty cycle implies that only a small fraction of SMBHs are active as UV-bright, luminous quasars at any given time, and it has relevant consequences in terms of the lifetime of high-z quasars, their obscuration fraction, and more generally the growth of SMBHs. We will explore this further in Sec. 3.5.2.

3.4.2 Characterizing the properties of [OIII] emitters

Our joint model for quasars and galaxies constrains the properties of these two populations simultaneously. As a result, all the properties that we have presented for quasars can also be studied for the high-z galaxy population. These are the galaxy luminosity-halo mass relation (Fig. 3.3, left panel), the GHMF (Fig. 3.3, right), and the galaxy duty cycle (Fig. 3.4, left panel). Before analyzing these quantities, we note that our model focuses only on [O III] emitters, as this sub-population of galaxies is the one that is targeted by the JWST NIRCam-WFSS observations from the EIGER survey. Therefore, all results that we will quote here refer to the properties of galaxies that are bright in the [O III] line; at these high redshifts, these galaxies are believed to be luminous, star-forming, and unobscured (e.g., Matthee et al. 2023).

The galaxy luminosity-halo mass relation (Fig. 3.3, left panel) is rather similar to the quasar luminosity-halo mass relation. The major differences can be found in the slope of this relation as well as in its normalization. The logarithmic slope of the galaxy luminosity-halo mass relation is shallower than the one concerning quasars, but steeper than linear ($\gamma^{(\mathrm{Gal})} \approx 2.3$). The normalization of this relation conspires with its slope to give an average galaxy luminosity at fixed halo mass that is brighter than the one of quasars at $\log_{10} M/\mathrm{M}_\odot \lesssim 11.5$, but dimmer at larger host halo masses[4]. This implies that, on average, quasars overshine galaxies at the high mass end of the HMF, while the opposite is true for the bulk of the halo population.

Nonetheless, if we look at the comparison between the QLF and the GLF in Fig. 3.2 (top panel), we see that our model predicts galaxies to be more abundant than quasars at all luminosities. This is because the scatter in the galaxy $L - M$ relation is rather large, and the duty cycle of galaxies is significantly larger than that of quasars (see below). Observationally, we know that the GLF drops below the QLF at luminosities around $\log_{10} L/\mathrm{erg\,s^{-1}} \approx 46$ (e.g., Bouwens et al. 2015; Matsuoka et al. 2018), so this implies that the extrapolation of the GLF at large luminosities based on our model is flawed. This is not a surprise, as here we assumed that a very simple power-law relation between galaxy luminosity and halo mass holds for the entire population of halos. This relation serves our purposes, as we want to match data for the GLF only in a rather narrow luminosity range, but it is probably too simplistic to capture the behaviour of the galaxy population at even larger luminosities/host masses.

Indeed, we know that the star formation efficiency is predicted to peak for halo masses of $\log_{10} M/\mathrm{M}_\odot \approx 11.5 - 12.5$, resulting in a break in the stellar mass-halo mass relation (e.g., Moster et al. 2013; Behroozi et al. 2019).

[4]Note, however, that the luminosity of galaxies only includes the flux emitted in the [O III] line, hence we expect the normalization of the galaxy $L - M$ relation to be higher when considering the total flux emitted from galaxies.

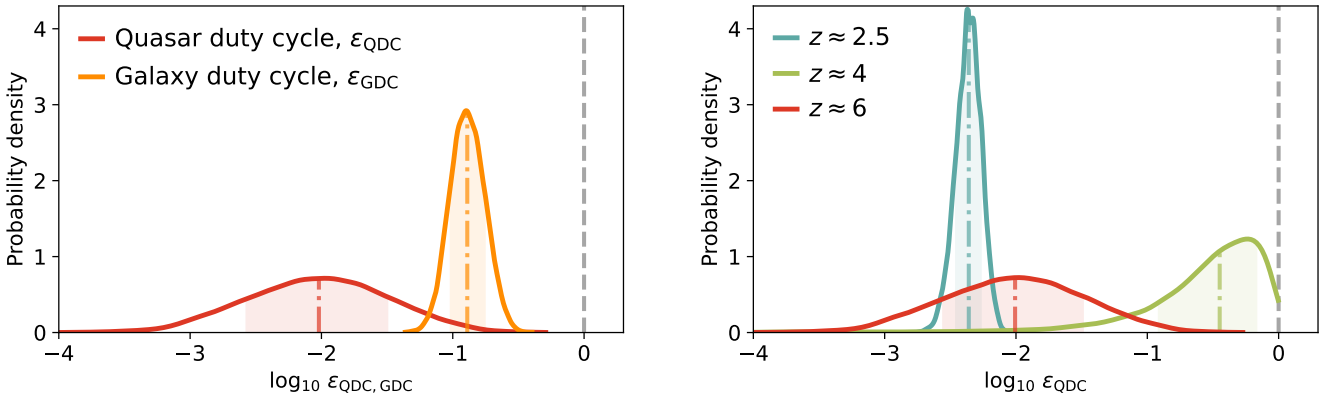

Figure 3.4: *Left*: Probability density function (PDF) for the quasar (red) and galaxy (orange) duty cycles at $z \approx 6$, obtained by randomly sampling the Markov chains for the posterior distribution 2000 times. The median and 1σ uncertainties for the PDFs are shown with a dot-dashed line and shaded areas, respectively. The dashed vertical line corresponds to the maximum possible value of the duty cycle, $\varepsilon_{\rm DC} = 100\%$. *Right*: Same as the left panel, but for the quasar duty cycles at different redshifts: $z \approx 2.5$ (blue), $z \approx 4$ (green), and $z \approx 6$ (red). Results at redshift lower than $z \approx 6$ are taken from P24.

While the luminosity range of the GLF data considered here is not large enough to constrain this break in the context of our model, we can see what would be the effect of a more physically motivated choice for the galaxy $L - M$ by making the arbitrary assumption that this relation flattens above $\log_{10} M/\mathrm{M}_\odot \approx 11.5$ (dashed line in the left panel of Fig. 3.3). In practice, we assume that the galaxy CLF in eq. 3.1 remains the same, but we vary the luminosity-mass relation on which it is based (eq. 3.2) by manually inserting a flattening above a threshold halo mass. We find that all the quantities but the GLF remain unchanged; the new median GLF is plotted with a dashed line in Fig. 3.2 (top panel). Indeed, we see that with this simple assumption, the predicted GLF drops below the QLF at roughly the observed luminosity. A more comprehensive quasar/galaxy population model – that is outside the scope of this paper – would include a larger set of galaxy observations to properly constrain the shape of the break in the galaxy $L - M$ relation. The simple argument adopted here, however, shows that our framework is well-suited to represent quasars and galaxies in the luminosity/mass ranges of the data we aim to reproduce (Table 3.1).

The GHMF is shown in the right panel of Figure 3.3 (orange line). Again, we find a broad distribution of host masses, with a median value of $\log_{10} M/\mathrm{M}_\odot \approx 10.9$ ($\log_{10} M/\mathrm{M}_\odot = 10.88^{+0.04}_{-0.03}$) and a 1σ range of ± 0.3. Determining the characteristic host halo masses for [O III] emitters is an important result that is made possible by the analysis presented here. This population of galaxies is a major protagonist in JWST campaigns to study the high-z Universe via slitless spectroscopy (Kashino et al. 2023; Oesch et al. 2023; Wang et al. 2023). For this reason, a thorough characterization of their properties is key. Overall, the characteristic host mass that we find for [O III] emitters agrees well with the one measured at the same redshifts using Lyman break galaxies (LBGs) in HST photometric campaigns (Barone-Nugent et al. 2014; Dalmasso et al. 2024). This result strengthens the conclusion – coming from abundance arguments (Matthee et al. 2023) – that [O III] emitters may trace star-forming regions in high-z galaxies in a way that is similar to Lyman-break-selected systems.

We note that the shape of the GHMF (Fig. 3.3, right panel) is affected by the minimum mass we assume in our model, i.e., $\log_{10} M_{\mathrm{min}}/\mathrm{M}_\odot = 10.5$ (see Sec. 3.2). In other words, in our population model, we assume that galaxies live only in halos larger than this threshold mass, and that the GHMF goes to zero for lower masses. This choice is made in the context of our framework because the FLAMINGO-10k simulation introduced in Sec. 3.2.2 cannot resolve halos with lower masses. There is no physical motivation, however, for this choice, as there could be a population of bright galaxies that are residing in lower-mass halos. In particular, we believe that extending the GHMF distribution to lower halo masses would bring the median value found here ($\log_{10} M/\mathrm{M}_\odot \approx 10.9$) down to slightly lower values. This is because the GHMF distribution is artificially skewed towards larger halo masses

because of the halo mass threshold imposed in our simulation: the halo mass corresponding to the peak of the GMHF distribution ($\log_{10} M/\mathrm{M}_\odot \approx 10.7$) is lower than the median ($\log_{10} M/\mathrm{M}_\odot \approx 10.9$). Indeed, a lower median value of $\log_{10} M/\mathrm{M}_\odot \approx 10.7$ is in closer agreement with the result found in E24, where the same simulation presented here was coupled with a "step-function" HOD model for quasars and galaxies. The authors found a *minimum* host mass for [O III] emitters of $\log_{10} M/\mathrm{M}_\odot \approx 10.56$, which can be translated into a median mass of $\log_{10} M/\mathrm{M}_\odot \approx 10.65$. Nonetheless, we believe that extending the model to lower halo masses would not significantly impact the conclusions presented here: we experimented with different prescriptions for the GHMF and always found similar results, with the median value of the GHMF of $\log_{10} M/\mathrm{M}_\odot \approx 10.8 - 10.9$) and the peak of the GHMF distribution at $\log_{10} M/\mathrm{M}_\odot \approx 10.6 - 10.7$. Using a simulation with a smaller volume and higher resolution, one could resolve halos down to much lower masses and hence fully capture the properties of galaxies and their host halos. However, this is not the goal of our work, as the primary focus of our analysis is the relation between quasars and the galaxies in their environments, which can only be captured with a large-volume simulation given the rarity of quasars at high-z.

The galaxy duty cycle, $\varepsilon_{\mathrm{GDC}}$, is a measure of how many halos host galaxies that can be observed in [O III] compared to the global halo population with the same characteristic masses. In our model, we infer a value for the galaxy duty cycle of $\varepsilon_{\mathrm{GDC}} = 12.9^{+4.7}_{-3.3}\%$. This is once again in agreement with the duty cycle values inferred from LBG clustering analysis (e.g., Dalmasso et al. 2024). We note here that the notion of "duty cycle" is primarily utilized in the context of quasars rather than galaxies, as gas accretion on SMBHs – that is believed to be associated with the triggering of quasar activity – is assumed to be episodic, and hence the whole process is cyclic in cosmic time. In the context of galaxies, it is probably easier to talk about an "occupation fraction" of [O III] emitters, implying that only a fraction of halos is hosting galaxies whose [O III] emissions are bright enough to be detectable and not obscured by dust. However, it is also relevant to point out that if [O III] emitters, as argued before, trace unobscured star formation, they may also be subject to rapid change in their luminosity as the star formation process is also thought to be episodic, especially at high redshifts (e.g., Faucher-Giguère 2018; Pallottini & Ferrara 2023). Indeed, UV-variability (e.g., Shen et al. 2023; Sun et al. 2023) has been argued to play a key role in explaining the over-abundance of bright galaxies that was indicated by JWST imaging at very high-z (e.g., Naidu et al. 2022; Finkelstein et al. 2024).

Our duty cycle measurement cannot determine the amount of variability in the galaxy lightcurves, as it only offers an integral constraint on the total light emitted (in the [O III] line) by star-forming galaxies over the entire history of the Universe. In other words, it is only sensitive to the zeroth moment of the galaxy's unobscured lightcurve distribution. Nonetheless, the

value of the duty cycle inferred here represents an important independent characterization of the star formation history of high-z galaxies, and it nicely complements probes of the burstiness of the high-z star formation process coming from spectral energy distribution (SED) fitting (e.g., Looser et al. 2023; Endsley et al. 2024; Cole et al. 2023).

Another interesting point to make here is that the duty cycle/occupation fraction that we measure for galaxies sets an upper limit on the contribution of obscured star formation to the total galaxy mass growth at early times. This is because our measurements tell us that $\gtrsim 15\%$ of $z \approx 6$ galaxies are [O III] -bright, and hence the fraction for which star formation is obscured by dust cannot be higher than $\approx 85\%$. This is an interesting constraint that can be directly compared with the estimated fraction of obscured star formation coming from ALMA observations (e.g., Algera et al. 2023). We will return to the point of obscuration in the context of quasars and SMBH growth in Sec. 3.5.2

3.5 Discussion

In the analysis performed above, we could successfully match the luminosity functions and the clustering properties of quasars and galaxies at $z \approx 6$ provided that: (a) there exist non-linear relations between quasar/galaxy luminosity and halo mass; (b) these relations have significant log-normal scatter ($\sigma \approx 0.5 - 1$ dex), and the one for quasars is steeper ($\gamma^{(\mathrm{QSO})} \approx 3.2$) than the one for galaxies ($\gamma^{(\mathrm{Gal})} \approx 2.3$); (c) following these relations, luminous quasars ($\log_{10} L/\mathrm{erg\,s^{-1}} \gtrsim 47$) are hosted by halos with mass $\log_{10} M/\mathrm{M_\odot} \approx 12.5$, while galaxies ($\log_{10} L/\mathrm{erg\,s^{-1}} \gtrsim 42.5$) are hosted by much smaller halos with $\log_{10} M/\mathrm{M_\odot} \approx 10.9$; (d) (UV-bright) quasars occupy only a small fraction of halos with a duty cycle $\varepsilon_{\mathrm{QDC}} \approx 0.9\%$, while the occupation fraction/duty cycle of galaxies is significantly larger, $\varepsilon_{\mathrm{GDC}} \approx 13\%$.

In the following, we further elaborate on this picture by focusing on the properties of high-z quasars, studying their implications for SMBH accretion and growth and their evolution with cosmic time.

3.5.1 Quasar properties across cosmic time

In P24, we applied a very similar framework to the one presented here to model the auto-correlation and luminosity functions of quasars at $z \approx 2.5$ and $z \approx 4$. As a result, we obtained the quasar luminosity-halo mass relation, the QHMF, and the quasar duty cycle at these two different redshifts, and discussed how the properties of quasars seem to evolve rapidly between these two epochs. Thanks to the analysis performed here, we can now extend this discussion to include the properties of $z \approx 6$ quasars, and attempt to paint a coherent picture of quasar evolution in the first few billion years of the

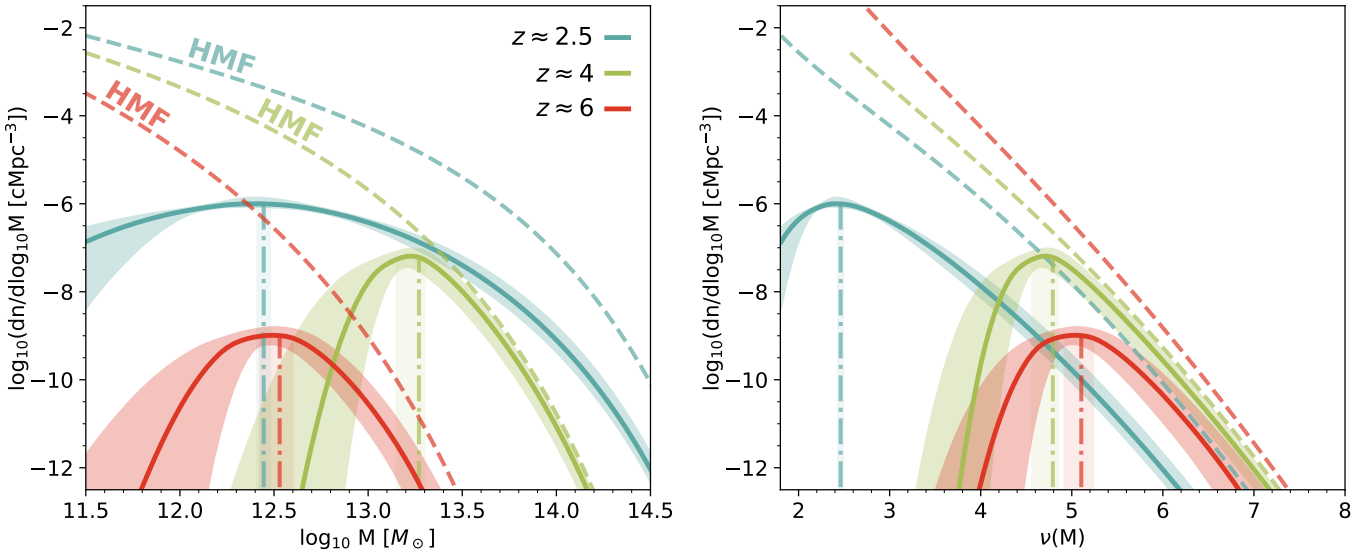

Figure 3.5: *Left*: Quasar-host mass function (QHMF) at $z \approx 2.5$ (solid blue line), $z \approx 4$ (solid green line), and $z \approx 6$ (solid red line) as a function of the halo mass, M. The halo mass functions (HMFs) at the same redshifts are shown with dashed lines and color-coded in the same way as the QHMFs. The dashed-dotted lines represent the median values of the QHMF distributions (see eq. 3.21), while shaded regions represent 1σ uncertainties on the various quantities. *Right*: Quasar-host mass functions (QHMFs; solid lines) and halo mass functions (HMFs; dashed lines) as a function of the peak height, $\nu(M)$, at different redshifts. Color codes and other quantities are the same as in the left panel.

Universe. The right panel of Fig. 3.4 shows the PDFs for the inferred values of the quasar duty cycles at $z \approx 2.5$ (blue), $z \approx 4$ (green), and $z \approx 6$ (red). The first two curves are obtained by sampling the posterior distributions for the parameters from P24 (see their Fig. 5), while the last one is the same as in the left panel. The same plot but for the QHMF is shown in the left panel of Fig. 3.5.

Quite interestingly, we see that the evolution of the QHMF and the quasar duty cycle with redshift do not follow a monotonic trend. The duty cycle is low ($\lesssim 0.5\%$) at $z \approx 2.5$, but it increases rapidly to values $\gtrsim 50\%$ at $z \approx 4$. At even higher redshifts, however, the duty cycle seems to drop again to $\lesssim 1\%$. Despite the relatively large uncertainty on our $z \approx 6$ measurement, the difference with the result obtained at $z \approx 4$ is rather remarkable (Fig. 3.4, right panel). An analogous trend with redshift can be observed by considering the median of the QHMF distribution, which represents the characteristic mass for the population of halos that are hosting bright quasars (Fig. 3.5, left panel): this median mass is $\approx 10^{12.4-12.5}\,\mathrm{M}_\odot$ for $z \approx 2.5$ and $z \approx 6$, while it grows to $\approx 10^{13.3}\,\mathrm{M}_\odot$ at $z \approx 4$.

As discussed in P24, the rather extreme values of the duty cycle and the host masses that we find at $z \approx 4$ are a consequence of the very strong quasar clustering measured by Shen et al. (2007). Using data from the Sloan Digital Sky Survey (SDSS), Shen et al. (2007) find a value of the quasar auto-correlation length, $r_{0,\mathrm{QQ}}$, of $\approx 24\,\mathrm{cMpc}\,h^{-1}$, which is significantly higher than the value $r_{0,\mathrm{QQ}} \approx 8\,\mathrm{cMpc}\,h^{-1}$ measured by Eftekharzadeh et al. (2015) (see also Ross et al. 2009; Shen et al. 2009; White et al. 2012) at $z \approx 2-3$ using the BOSS survey. The strong quasar clustering at $z \approx 4$, combined with a rather large abundance of bright quasars at the same redshift ($\approx 3 \times 10^8\,\mathrm{cMpc}^{-3}$), implies that only very massive halos can host active SMBHs and a large fraction of them are continuously shining as quasars at any given moment (i.e., the quasar duty cycle is large). This – as discussed by several works (P24,White et al. 2008; Shankar et al. 2010b) – is only possible provided that the scatter in the relation between quasar luminosity and halo mass is very low ($\sigma \lesssim 0.3$ dex).

The analysis presented here shows that the trend hinted by the Shen et al. (2007) quasar clustering measurements at $z \approx 4$ does not seem to extend further to higher redshifts. Using the data from E24, we have shown that the characteristic host mass of quasars at $z \approx 6$ is not as large, and only a small fraction of these SMBH-hosting halos are actually shining as bright quasars at any given time. As a consequence, the tight constraints on the scatter between quasar luminosity and halo mass are not in place at $z \approx 6$, and our model finds a larger value for the scatter of $\sigma \approx 0.6$ dex, although lower values are also compatible with the data (Fig. 3.1). Overall, these results may suggest that the measurements of quasar clustering at $z \approx 4$ (Shen et al. 2007) may be overestimated (see also He et al. 2018; de Beer et al. 2023), and that the constraints on the host masses, quasar

duty cycle, and scatter in the $L - M$ relation at $z \approx 4$ may need to be relaxed to some extent. If that is the case, our results at $z \approx 6$ suggest that quasars are hosted, on average, by a small fraction of the population of halos with masses in the range $\approx 10^{12} - 10^{13}\,\mathrm{M}_{\odot}$, in line with the situation at $z \approx 2 - 3$. This result may favor a picture in which there exists a range of halo masses for which quasar activity can be supported that is independent of cosmic time. According to this picture, halos whose masses are lower than this range cannot be responsible for a significant fraction of the black holes that are capable of turning into bright quasars, while for very massive hosts ($\log_{10} M/\mathrm{M}_{\odot} \approx 13$) quasar activity is quenched by feedback mechanisms (e.g., Hopkins et al. 2007b; Fanidakis et al. 2013; Caplar et al. 2015).

On the other hand, the measurements from Shen et al. (2007) appear to be solid, being based on a large (≈ 5000) spectroscopic sample of high-z quasars from SDSS, and they are also backed up by estimates of the small-scale quasar clustering inferred from independent samples of $z \approx 4-5$ binary quasars (Hennawi et al. 2010; McGreer et al. 2016; Yue et al. 2021). It is thus worth taking the Shen et al. (2007) clustering data at face value, and exploring the implications of their results in terms of the evolution of quasar properties at early cosmic times. The Shen et al. (2007) measurements suggest that at high redshifts quasar activity tends to take place only in the most massive halos, tracking halo growth across cosmic time (Hopkins et al. 2007b). It is interesting to note that our $z \approx 6$ results do not necessarily disfavor this scenario. In fact, our inferred QHMFs suggest that quasars live *in equivalent halos* at $z \approx 4$ and $z \approx 6$, while they live in very different environments at lower redshifts. This can be understood by looking at the right panel of Fig. 3.5, which shows the QHMFs at different redshifts plotted as a function of the peak height, $\nu(M)$. The peak height is defined as $\nu(M, z) = \delta_c/\sigma(M, z)$ – with $\delta_c \approx 1.69$ being the critical linear density for spherical collapse and $\sigma^2(M, z)$ the variance of the linear density field smoothed on a scale $R(M)$ [5]. It is a way to relate the masses of halos at any redshifts to the strength of the fluctuations in the initial conditions of the original linear density field. Therefore, large (small) peak heights correspond to very over(under)-dense environments, independently of redshift.

The right panel of Fig. 3.5 shows that quasars tend to be hosted by very rare $\approx 5\sigma$ fluctuations both at $z \approx 6$ and $z \approx 4$. This suggests that the same kind of rare and very biased halos host bright quasars at early cosmic times,

[5] We compute $\nu(M, z)$ using the python package COLOSSUS (Diemer 2018) and setting the same cosmology as the FLAMINGO-10k simulation (Sec. 3.2.2). However, we mention the caveat that the definition of peak heights implicitly assumes that halo masses are based on the spherical overdensity formalism, and it only applies to the current masses of central halos (and not to satellites). In our analysis (Sec. 3.2.2.2), we assume a halo mass definition based on peak bound masses instead, and include the contribution of satellites as well. Nonetheless, we believe that the effects of the differences in halo mass definition are relatively small and that the final values we obtain for the peak heights are not impacted significantly by these factors.

and that these host halos are more massive at $z \approx 4$ than at $z \approx 6$ only because they grow via mergers/accretion during the ≈ 700 million years of cosmic time that separate these two redshifts. In the lower redshift Universe ($z \approx 0 - 3$), instead, the situation is quite different, with quasars being hosted by a new, less biased population of halos which corresponds to $\lesssim 3\sigma$ fluctuations in the density field.

In this scenario, the key difference between $z \approx 6$ and $z \approx 4$ is the *duty cycle*: while at $z \approx 4$ almost all of the most massive halos need to host UV-bright quasars, the fraction of these same halos that are revealed as quasar hosts at $z \approx 6$ is dramatically smaller. This could be caused by either much shorter and sparser accretion episodes at very early cosmic times or a much larger obscuration fraction characterizing early SMBH accretion. It is of great interest to relate these arguments to our current paradigm of SMBH growth: this will be the subject of Sec. 3.5.2.

In order to discriminate between the scenarios discussed here and to paint a complete evolution of quasar activity across cosmic time, it is essential to investigate the clustering of quasars at high redshifts with new methods and new observational campaigns. In this sense, the next few years promise to bring a new wealth of data with the combined action of JWST mapping quasar-galaxy clustering at different redshifts using NIRCam WFSS (Kashino et al. 2023; Wang et al. 2023), and the DESI survey (DESI Collaboration et al. 2016) using ground-based spectroscopy to unveil a new, large sample of quasars up to $z \lesssim 5$ that can be used to compute the quasar auto-correlation function with a much higher precision.

We conclude by mentioning the caveat that the QHMFs shown in Fig. 3.5 are obtained by setting luminosity thresholds that vary according to the ones used in observational data. In other words, the definition of "bright" quasars we employ is redshift-dependent, and it is based on the depth of the survey that was used for the clustering measurements. In Appendix 3.D, we show the same QHMFs obtained by setting a uniform luminosity threshold of $\log_{10} L_{\rm thr}/{\rm erg\,s^{-1}} = 46.7$, which is the same luminosity threshold as used by Shen et al. (2007) at $z \approx 4$ and roughly corresponds to the break of the quasar luminosity function at all redshifts $z \gtrsim 2$ (e.g., Khaire & Srianand 2015; Kulkarni et al. 2019). The resulting QHMF shifts towards higher (lower) halos masses at $z \approx 2.5$ ($z \approx 6$), due to the different luminosity thresholds employed in observations with respect to the one at $z \approx 4$. Nonetheless, the global picture that we presented in this Section remains unchanged: quasars seem to be hosted by $\log_{10} M/{\rm M}_\odot \gtrsim 13 - 13.5$ halos only at $z \approx 4$, but when relating halo masses to their large-scale environments by using the peak height formalism, we find a direct connection between $z \approx 4$ and $z \approx 6$ and a divergent behavior at lower redshifts.

3.5.2 The quasar duty cycle and SMBH growth

One of the key results of our analysis is that the quasar duty cycle we obtain at $z \approx 6$ is rather low ($\approx 0.9\%$), in stark contrast with the very high one ($\gtrsim 50\%$) measured at $z \approx 4$ from the Shen et al. (2007) data (Fig. 3.4, right panel). As detailed in, e.g., P24, these duty cycles can be directly converted into estimates of the total time SMBHs shine as bright quasars (i.e., the integrated quasar lifetime, $t_{\rm Q}$) via the simple relation $t_{\rm Q} = t_{\rm U}(z)\ \varepsilon_{\rm DC}$ – with $t_{\rm U}(z)$ being the age of the Universe at a given redshift. Using the values of the duty cycles mentioned above, we obtain $t_{\rm Q} \approx 0.1 - 1$ Gyr at $z \approx 4$, and a smaller $t_{\rm Q} \approx 10$ Myr at $z \approx 6$. It is important to investigate the discrepancy between the values obtained at these two redshifts further, as the study of the timescales of quasar activity at high redshift is intrinsically connected with the formation and evolution of SMBHs in the Universe.

As discussed in the Introduction, our current paradigm of SMBH growth is founded on the idea that SMBHs grow by accretion, and that a small fraction of the accreted rest mass is converted into radiation and gives rise to the quasar phenomenon. According to this paradigm, the growth of SMBHs is always concomitant with the formation of a bright quasar. For this reason, the total time a SMBH shines as a quasar (i.e., the quasar lifetime) is related to the total mass that has been accreted onto the SMBH. This argument has been proposed in many different variations in the past (e.g., Soltan 1982; Martini & Weinberg 2001; Yu & Tremaine 2002), and it represents one of the cornerstones of our understanding of quasar/SMBH evolution.

At high redshift ($z \gtrsim 6$), the connection between the quasar lifetime and SMBH growth is even more relevant due to the limited amount of cosmic time ($\lesssim 1$ Gyr) that is available to grow black holes to the observed masses of $\approx 10^{8-9}\,{\rm M}_{\odot}$ (Fan et al. 2023). Assuming Eddington-limited growth with a standard radiative efficiency of $\approx 10\%$, one finds that only by postulating $t_{\rm Q} \sim 0.1 - 1$ Gyr (i.e., a quasar duty cycle $\gtrsim 10\%$) it is possible to explain the presence of such black holes in the early Universe starting from massive black hole seeds of $\approx 10^{3-5}\,{\rm M}_{\odot}$ (e.g., Inayoshi et al. 2020; Pacucci & Loeb 2022). This argument agrees well with the long lifetime inferred by our model at $z \approx 4$ (see P24 for further discussion), but it is in plain tension with the low duty cycle at $z \approx 6$ that we inferred in this work.

This tension between the long timescales required by SMBH growth and the short timescales that seem to be associated with high-z quasar activity has already been investigated in the context of quasar proximity zones and damping wing features. By looking at quasar rest-frame UV spectra, several studies at $z \approx 4 - 7$ have argued that the inferred quasar lifetimes range between $t_{\rm Q} \approx 0.1 - 10$ Myr (e.g., Khrykin et al. 2016, 2019; Eilers et al. 2018, 2020; Davies et al. 2018, 2019, 2020; Worseck et al. 2016, 2021; Ďurovčíková et al. 2024), and do not seem to reach the very large values required by SMBH growth models. Constraints based on proximity zones/damping

wings are sensitive to the local conditions of each quasar environment and only probe a fraction of the past quasar lightcurve, so the direct connection between these results and the ones related to quasar clustering – which probe the global population of quasars and can only constrain their total lifetime – is non-trivial in the presence of rapidly varying and/or flickering lightcurves (e.g., Davies et al. 2020; Satyavolu et al. 2023).

Nonetheless, the cumulative evidence coming from these very different probes of quasar activity indicates that our standard paradigm for SMBH growth at high z may need to be thoroughly reconsidered: not only is there very little cosmic time to grow the SMBHs to the billion solar masses that we observe for bright $z \approx 6-8$ quasars, we also lack evidence for this accretion taking place in the form of UV-bright quasar emission at $z \gtrsim 6$. Proposed solutions to this problem include a very low radiative efficiency $\lesssim 0.1-1\%$ – which implies that only a very small fraction of the accreted mass is converted into quasar light – or a very large population of obscured SMBHs at high-z that are not visible as UV-bright quasars but continue to grow actively at all epochs (e.g., Davies et al. 2019). This latter hypothesis is particularly relevant, as a large obscured fraction for $z \gtrsim 6$ quasars has been proposed both in the context of cosmological simulations (e.g., Ni et al. 2020; Vito et al. 2022; Bennett et al. 2024) and multi-wavelength observations (Vito et al. 2018; Circosta et al. 2019; D'Amato et al. 2020; Gilli et al. 2022; Endsley et al. 2024). Recently, JWST data have unveiled a new population of candidate dust-obscured active galactic nuclei (AGN) that can only be found at high redshifts (Harikane et al. 2023; Matthee et al. 2024b; Kocevski et al. 2023; Maiolino et al. 2024; Greene et al. 2024; Kokorev et al. 2023, 2024a; Lin et al. 2024), and may suggest a rapid evolution of the obscuration properties of AGN/quasars in the early Universe.

3.6 Summary

In this work, we have modeled the demographic and clustering properties of quasars (i.e., type-I, UV-bright systems) and galaxies (i.e., [O III] emitters) at $z \approx 6$ using an extension of the framework introduced in Pizzati et al. (2024a) (P24; see their Figure 1). The model presented here builds on a new, state-of-the-art N-body simulation from the FLAMINGO suite (Schaye et al. 2023) (the "FLAMINGO-10k" run) that has the same resolution as the original FLAMINGO DMO high-resolution run (CDM particle mass of $8.40 \times 10^8\,\mathrm{M}_\odot$) but a much larger volume ($L = 2.8\,\mathrm{cGpc}$).

Thanks to this simulation, we can model the properties of $z \approx 6$ quasars and galaxies simultaneously; these include (Table 3.1): the galaxy luminosity function (Matthee et al. 2023), the quasar luminosity function (Schindler et al. 2023), the quasar-galaxy cross-correlation function and the galaxy

auto-correlation function (Eilers et al. 2024), and the quasar auto-correlation function (Arita et al. 2023, considered separately in Appendix 3.C).

The model we employ is founded on a Conditional Luminosity Function (CLF) framework. We assume a CLF for both quasars and galaxies, with identical parameterizations, i.e., power-law relations between quasar/galaxy luminosity and halo mass ($L \propto M^{\gamma}$) with log-normal scatter σ. We also include an active fraction, $f_{\rm on}$, to account for the fraction of quasars/galaxies that are too dim or not active and hence cannot be detected by observations.

The CLFs effectively link the population of halos in the simulated volume to the ones of quasars/galaxies. Therefore, once the halo mass function is known, we can directly obtain the quasar/galaxy luminosity function and the quasar-/galaxy-host mass function (QHMF/GHMF). The QHMF/GHMF can be coupled to the cross-correlation functions of halos with different masses to model the clustering properties (auto-/cross-correlations) of quasars and galaxies simultaneously.

As detailed in P24, the halo mass function and the cross-correlation functions of halos with different masses are extracted from the simulation and used to construct analytical fitting functions. We stress the fact that the framework introduced here is general, and can be used to predict the clustering and/or demographic properties of any populations of halo tracers (see also Appendix 3.A).

We summarise below the main findings of our analysis:

- We jointly model all the observational data in Table 3.1 except for the quasar auto-correlation function (Arita et al. 2023), which we analyze separately in Appendix 3.C. We find a very good match between our predictions and observations for all the quantities considered (Fig. 3.2). The posterior distribution for the model parameters favors relatively large values for the scatter both in the quasar luminosity-halo mass relation and in the galaxy luminosity-halo mass relation ($\sigma \approx 0.6 - 0.8\,$dex), with the relation for quasars being steeper than the one for galaxies (Fig. 3.3, left panel). The active fraction, on the other hand, is larger for galaxies ($f_{\rm on} \approx 25\%$) than for quasars ($\approx 4\%$). Interestingly, the luminosity-halo mass relations inferred in Fig. 3.3 (left) imply that galaxies outshine quasars (i.e, the average [O III] luminosity of galaxies is larger than the bolometric luminosity of quasars) at halo masses of $\log_{10} M/{\rm M}_{\odot} \lesssim 11.5$.

- According to the results above, $z \approx 6$ quasars live on average in $\approx 10^{12.5}\,{\rm M}_{\odot}$ halos, with a mass distribution that is quite broad, from $\approx 10^{12.1}\,{\rm M}_{\odot}$ halos to $\approx 10^{12.8}\,{\rm M}_{\odot}$ (according to the 2σ limits of the QHMF distribution; see right panel of Fig. 3.3). This broad QHMF distribution implies that quasars inhabit rather diverse environments at high-z. This, together with the contribution of cosmic variance, may explain the large quasar-to-quasar variance in terms of environments

that was reported by Eilers et al. (2024), as well as the contradictory claims that have been made based on previous observations (e.g., Kim et al. 2009; Mazzucchelli et al. 2017b; Mignoli et al. 2020).

- Despite the rather large uncertainties, we are able to constrain the $z \approx 6$ (UV-bright) quasar duty cycle to $\varepsilon_{\rm QDC} \lesssim 1\%$ (Fig. 3.4, left panel). This relatively low value translates to quasar lifetimes of $\approx 10\,{\rm Myr}$, in stark contrast with the very long lifetimes required at high z by the standard picture of SMBH formation and growth (e.g., Inayoshi et al. 2020). This finding challenges our paradigm for SMBH growth at high-z, and suggests that most of the black hole mass growth may have happened in highly obscured and/or radiatively inefficient environments (see also Davies et al. 2019).

- As expected, the properties of galaxies (i.e., [O III] emitters) that we obtain are rather different from the ones of quasars (Fig. 3.2-3.4). The characteristic host mass for [O III] emitters that we measure from the GHMF is $\approx 10^{10.9}\,{\rm M}_{\odot}$, in line with the one estimated from LBG clustering measurements (e.g. Barone-Nugent et al. 2014; Dalmasso et al. 2024). This suggests that [O III] emitters may be tracing the population of high-z actively star-forming galaxies in a way that is similar to what LBGs have been doing in the Hubble Space Telescope (HST) era. The galaxy duty cycle that we infer is larger than the one of quasars, $\varepsilon_{\rm GDC} \approx 13\%$, suggesting that a significant fraction of high-z galaxies are UV-bright and actively star-forming at $z \approx 6$. This sets an implicit constraint on the fraction of galaxies that are quenched and/or obscured at the same redshifts.

- By comparing the properties of quasars at $z \approx 6$ obtained in this work with the ones discussed in P24 for $z \approx 2.5$ and $z \approx 4$, we find that the evolution of these properties with redshift seems to follow a non-monotonic trend (Fig. 3.5). The characteristic quasar-host mass and the quasar duty cycle have similar values at $z \approx 2.5$ and $z \approx 6$, but they increase to significantly higher values at $z \approx 4$ due to the strong quasar clustering measured by Shen et al. (2007). We discuss whether the conjunction between $z \approx 2.5$ and $z \approx 6$ may suggest that quasar properties are more or less stable across cosmic time, which would imply that the $z \approx 4$ quasar clustering measurements are overestimated. We also present a picture, however, in which the bulk of quasar activity takes place in very rare and over-dense environments (corresponding to $\approx 5\sigma$ fluctuations in the initial linear density field) at $z \approx 4$ and $z \approx 6$, while it migrates to a larger population of less biased halos at lower-z. Further observational work is needed to distinguish between these scenarios and map the evolution of quasar properties across cosmic time.

The analysis presented in this paper lays down a simple but powerful framework that exploits observations to characterize the properties of SMBHs and galaxies in the early Universe. New data and more detailed modeling can improve the constraints that we get in the context of this framework significantly.

Observationally, the ASPIRE survey (Wang et al. 2023) will soon complement observations from EIGER (Kashino et al. 2023; Eilers et al. 2024) by measuring the cross-correlation function for a larger sample of 25 moderately luminous quasars at $z \approx 6.5 - 6.8$. The enlarged sample provided by ASPIRE will be extremely useful for reducing the uncertainties in our model parameters as well as for quantifying the quasar-to-quasar variance in the cross-correlation function. In the near future, new observations from JWST could complement the ASPIRE and EIGER surveys by determining the clustering properties of quasars and galaxies in a wider redshift range as well as for the faint end of the quasar luminosity function.

In parallel with the acquisition of new observational data, the model presented here could be developed further to study the variance of the measured correlation function theoretically, and could be extended to take into account the velocity information coming from direct measurements of the redshift-space correlation function (e.g., Costa 2024). Results at different redshifts could also be linked together by developing an evolutionary model following the growth of supermassive black holes and the evolution of quasar activity across cosmic time.

3.A Appendix: Details on the conditional luminosity function framework

Given any population of "tracer" ("T") objects that are hosted by dark matter halos and are visible in some electromagnetic band, we can write down their 2-d distribution in the tracer luminosity-host halo mass plane, $n(L, M)$, as:

$$n(L, M) = \mathrm{CLF}(L|M)\, n_{\mathrm{HMF}}(M), \tag{3.17}$$

where $n_{\mathrm{HMF}}(M)$ is the halo mass function. The quantity $\mathrm{CLF}(L|M)$ is known as the conditional luminosity function, and it links in a statistical sense the population of dark matter halos to the population of tracer objects (e.g., Yang et al. 2003; Ballantyne 2017a,b; Bhowmick et al. 2019; Ren et al. 2020).

In this framework, we assume that every halo between a minimum mass M_{min} and a maximum mass M_{max} hosts a tracer object[6]. The luminosity L of this tracer can be defined arbitrarily, but it has to depend solely on the

[6] M_{min} and M_{max} are chosen here according to the mass range that can be reliably modeled based on the cosmological simulation employed (see Sec. 3.2.2).

mass of the halo. Following these assumptions, a simple marginalization of $n(L, M)$ over halo mass gives the luminosity function of the tracer species, $n_{\rm TLF}$:

$$n_{\rm TLF}(L) = \int_{M_{\rm min}}^{M_{\rm max}} {\rm CLF}(L|M)\, n_{\rm HMF}(M)\, {\rm d}M. \tag{3.18}$$

Analogously, integrating over the luminosity dimension returns the distribution in mass of the tracers. If we include only objects above some threshold luminosity (set e.g. by the flux limit of observations), we can obtain a mass distribution for halos whose tracer object is brighter than $L_{\rm thr}$, $n_{\rm THMF}$:

$$n_{\rm THMF}(M|L > L_{\rm thr}) = n_{\rm HMF}(M) \int_{L_{\rm thr}}^{\infty} {\rm CLF}(L|M)\, {\rm d}L. \tag{3.19}$$

Likewise, the aggregate probability for a halo of mass M to host a tracer with a luminosity above $L_{\rm thr}$ (also known as a Halo Occupation Distribution, HOD; see e.g., Berlind & Weinberg 2002) is:

$${\rm HOD}(M) = \frac{n_{\rm THMF}(M|L > L_{\rm thr})}{n_{\rm HMF}(M)} = \int_{L_{\rm thr}}^{\infty} {\rm CLF}(L|M)\, {\rm d}L. \tag{3.20}$$

Following, e.g., P24 (see also Ren et al. 2020), we can define the duty cycle of tracers above the luminosity threshold, $\varepsilon_{\rm DC}$, as the weighted average of the HOD above a threshold mass that is given by the median of the tracer-host mass function, $n_{\rm THMF}(M|L > L_{\rm thr})$. In other words, if we define the median of the $n_{\rm THMF}(M|L > L_{\rm thr})$ as the mass $M_{\rm med}$ satisfying the relation:

$$\int_{M_{\rm med}}^{M_{\rm max}} n_{\rm THMF}(M) = 0.5 \int_{M_{\rm med}}^{M_{\rm max}} n_{\rm THMF}(M), \tag{3.21}$$

then $\varepsilon_{\rm DC}$ can be expressed as:

$$\begin{aligned} \varepsilon_{\rm DC} &= \frac{\int_{M_{\rm med}}^{M_{\rm max}} n_{\rm HMF}(M)\, {\rm HOD}(M)\, {\rm d}M}{\int_{M_{\rm med}}^{M_{\rm max}} n_{\rm HMF}(M)\, {\rm d}M} = \\ &= \frac{\int_{M_{\rm med}}^{M_{\rm max}} n_{\rm THMF}(M|L > L_{\rm thr})\, {\rm d}M}{\int_{M_{\rm med}}^{M_{\rm max}} n_{\rm HMF}(M)\, {\rm d}M}. \end{aligned} \tag{3.22}$$

These relations hold for any tracer populations that satisfy the assumptions made above. In P24, we have considered SMBHs as tracer objects, assuming that every halo hosts a SMBH at its center emitting at some luminosity L. If the luminosity L is high enough, the SMBH becomes an

active quasar, and so we can use the conditional luminosity framework to obtain the quasar luminosity function ($n_{\rm QLF}$; analogous to eq. 3.18), the quasar-host mass function ($n_{\rm QHMF}$; analogous to eq. 3.19), and the quasar duty cycle ($\varepsilon_{\rm QDC}$; analogous to 3.22).

As commonly assumed in the literature (e.g., Yang et al. 2003; van den Bosch et al. 2003), galaxies are also tracers of the dark matter halo distribution. Following the P24 approach, we can then assume a conditional luminosity function for galaxies, and adapt the relations above to obtain the galaxy luminosity function ($n_{\rm GLF}$; analogous to eq. 3.18), the galaxy-host mass function ($n_{\rm GHMF}$; analogous to eq. 3.19), and the galaxy duty cycle ($\varepsilon_{\rm GDC}$; analogous to 3.22).

In Sec. 3.2.1, we write down explicitly the quasar/galaxy conditional luminosity functions adopted in this work[7], and provide more details on how to connect the quantities defined here to observations.

3.B Appendix: Results for the fitting of the halo cross-correlation functions

As described in Sec. 3.2.2, we compute the cross-correlation functions between $z \approx 6$ halos in different mass bins, $\xi_h(M_j, M_k; r)$, and then fit the results with a suitable parametrization of the radial and mass dependences. The details of the fitting are summarized in the main text and described at length in P24. Here, we focus on the results of these fits, comparing them to the actual correlation functions computed numerically from simulations and discussing their validity in the context of the problem we are facing here.

Figure 3.6 displays the overall results of the fit. The first two rows display the resulting fitting function ($\rho_{\rm fit}(\nu(M_j), \nu(M_k), r) = \xi_h(M_j, M_k; r)/\xi_{\rm ref}(r)$, where $\xi_{\rm ref}(r)$ is a reference correlation function, see main text for details). Each panel in these rows show the values of $\rho_{\rm fit}(\nu(M_j), \nu(M_k), \bar{r})$ as a function of the two masses M_j and M_k at a different scale $\bar{r}$. The last two rows show the relative difference $(\rho/\rho_{\rm fit} - 1)$ between our fit and the values of $\rho(\nu(M_j), \nu(M_k), r) = \xi_h(M_j, M_k; r)/\xi_{\rm ref}(r)$ obtained from the simulation. According to these figures, our simple analytical framework can describe the behavior of cross-correlation functions for a wide range of masses and scales with a good degree of accuracy ($\lesssim 5 - 10\%$). This level of accuracy is sufficient for the data we aim to reproduce here, as both the auto-correlation function of quasars and the quasar-galaxy cross-correlation functions are only known at the $30\% - 100\%$ level. The most constrained quantity is the galaxy auto-correlation function, which is however still uncertain at more than $\gtrsim 10\%$ (Sec. 3.3.1).

[7] Note that these functions depend on the specific population of "quasars" and "galaxies" we model, as well as on the definition of their luminosity, L. We refer to Sec. 3.2.1 for more details.

The only notable exception for which our fit doesn't perform well is the case of high masses ($\log_{10} M_{j,k}/\mathrm{M}_\odot \approx 11.5$) and small scales ($r \lesssim 0.5\,\mathrm{cMpc}$). However, this behavior is expected as high-mass halos are quite rare, and hence the measured correlation functions suffer in general from significant shot noise. At small scales this is worsened by the fact that the correlation function is dominated by the clustering of satellite halos, which are in general less massive than $\log_{10} M/\mathrm{M}_\odot \approx 11 - 12$. As a result, the cross-correlation functions of very massive systems drop at $r \lesssim 0.5\,\mathrm{cMpc}$ because of halo exclusion. Our fit hinges upon a smooth dependence of the correlation functions on mass and radius, and it is not able to capture halo exclusion properly. Nonetheless, this is not an issue for our analysis, because the data we aim to fit do not probe this specific regime: the auto-correlation function of quasars from Arita et al. (2023) is only measured at very large scales ($r \gtrsim 40\,\mathrm{cMpc}$), while the quasar-galaxy cross-correlation function and the auto-correlation function of galaxies from E24 are dominated by the contribution of galaxies, which live in relatively low mass halos ($\log_{10} M/\mathrm{M}_\odot \approx 10.5 - 11$; see Section 3.4).

Figure 3.7 shows two more comparisons between the cross-correlation functions extracted from the simulation and our fitting functions. In the left panel, we show the cross-correlation terms $\xi_h(M, \tilde{M}, r)$ as a function of radius, for different values of the mass M. The mass $\tilde{M}$ is chosen to represent the bin $\log_{10} M/\mathrm{M}_\odot = 10.5 - 10.75$. Errors on the values extracted from simulations are Poissonian (Sec. 3.2.2). Note that to properly reproduce the correlations measured in simulations, we select halos in each mass bin and weigh the fitting functions according to the mass distribution of halos (i.e., the halo mass function). In this way, we can take into account the actual distribution of halo masses in our fitting framework. Overall, we confirm that the fits and the values from simulations agree at the $\approx 5 - 10\%$ level, with the expected exception of the most inner bin.

The right panel of Figure 3.7 shows the halo auto-correlation functions for each mass bin, $\xi_h(M, M, r)$. As already mentioned above, we note that the accordance between fits and simulations is again satisfactory with the notable exceptions of large halo masses – for which halos are rare and the measured correlation functions are noisy – and small scales – for which halo exclusion plays an important role and our fit is not able to capture it properly. Overall, this visual comparison between simulations and fits confirms the fact that our framework can properly reproduce cross-correlation functions at all scales, as well as auto-correlation functions, with the exception of the high mass bins at small scales.

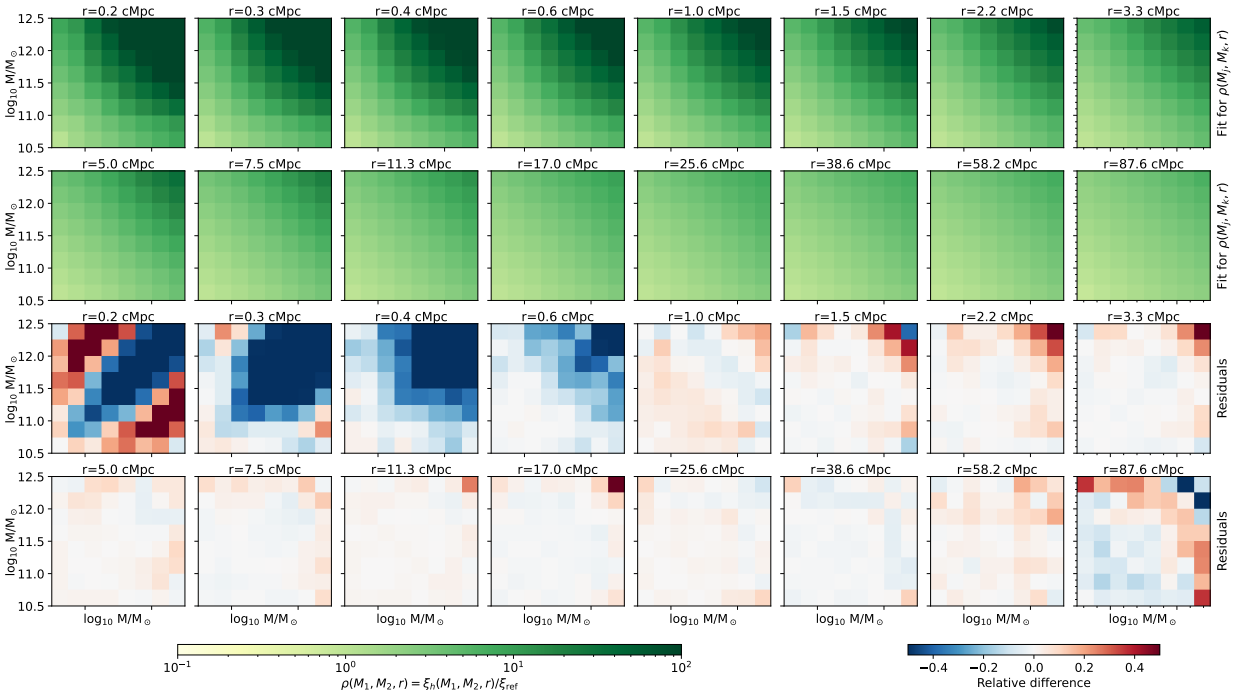

Figure 3.6: Results for the fitting of the $z \approx 6$ cross-correlation terms $\rho(M_j, M_k, r) = \xi_h(M_j, M_k; r)/\xi_{\rm ref}(r)$ (see Appedix 3.B for definitions). The two top rows show the fitting function $\rho_{\rm fit}(M_j, M_k, r)$ as a function of the two masses M_1 and M_2 for different values of the distance r. The last two rows show the relative difference between the fits and the values extracted from simulation.

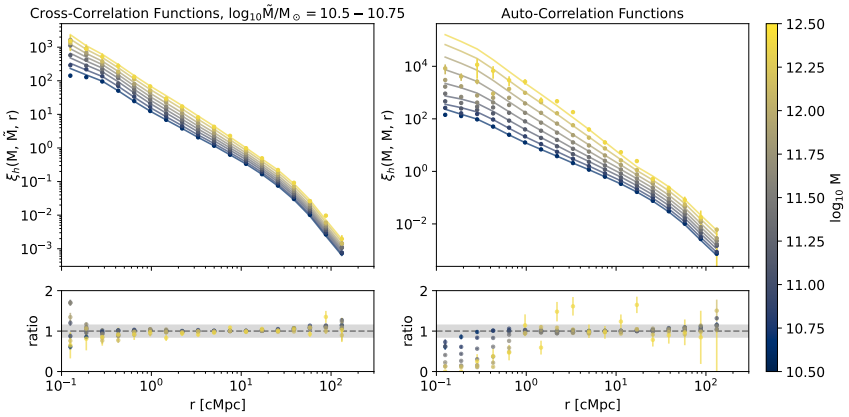

Figure 3.7: *Left:* Cross-correlation functions of halos in different mass bins, $\xi_h(M, \tilde{M}, r)$, at $z \approx 6$. The mass $\tilde{M}$ is set to correspond to the $\log_{10} M/\mathrm{M}_\odot = 10.5 - 10.75$ bin, while the other mass is varied according to the color scale. Values extracted from simulations are shown as data points, with error bars given by the Poissonian statistics of pair counting (see Sec. 3.2.2). Solid lines represent the fitting functions to these simulated values. Relative differences between the fit and the simulation are shown in the bottom panel. *Right:* Same as the left panel, but for the auto-correlation functions of halos in different mass bins, $\xi_h(M, M, r)$.

3.C Appendix: Interpreting the auto-correlation measurements of $z \approx 6$ quasars

In this Section, we analyze the data concerning the quasar auto-correlation function from Arita et al. (2023). As detailed in Sec. 3.3.1, we decided to leave this dataset out of the joint fit performed in the main analysis because we realized that its constraining power was less strong than expected. In particular, we found that using only the Arita et al. (2023) data, we were not able to place significant constraints on any of our model parameters.

For this reason, we use here a much simpler model that should make the interpretation of the data straightforward. In particular, we choose to parameterize the quasar-host mass function (QHMF) in the following way:

$$n_{\mathrm{QHMF}}(M) = \varepsilon_{\mathrm{DC}} n_{\mathrm{HMF}}(M) \Theta(\log_{10} M - \log_{10} M_{\mathrm{min}}), \qquad (3.23)$$

with $\varepsilon_{\mathrm{QDC}}$ being the duty cycle and Θ the Heaviside step function. In practice, we assume a simple "step-function" halo occupation distribution (HOD) model, depending only on one single parameter, the minimum host mass, M_{min} (the duty cycle $\varepsilon_{\mathrm{QDC}}$ is completely irrelevant for clustering measurements).

For every value of M_{min}, we can take the resulting QHMF and use it to compute the quasar auto-correlation function, $\xi_{\mathrm{QQ}}(r)$, according to eq.

3.6. With a simple integration along the radial direction (eq. 3.9), we can then obtain the projected auto-correlation function, $w_{p,\mathrm{QQ}(r_p)}$, which can be compared directly with the Arita et al. (2023) data.

As detailed in Sec. 3.2.2.3, our model for the correlation functions consists of two components: a fit to simulations, $\xi_{h,\mathrm{fit}}$, and a prediction based on the linear halo bias formalism, $\xi_{h,\mathrm{lin}}$ (eq. 3.14). The former is used to model the small-scale clustering ($r \lesssim 20\,\mathrm{cMpc}$), while the latter is used to regularize the behaviour of simulations at large scales ($r \gtrsim 20\,\mathrm{cMpc}$). The key point, here, is that the Arita et al. (2023) data we aim to interpret cover only very large scales, with the innermost bin at $r \approx 40\,\mathrm{cMpc}$. For this reason, we can safely assume that our model is entirely in the linear theory regime, and assume $\xi_h = \xi_{h,\mathrm{lin}}$. In other words, the model we discuss in this context is not unique to our simulations; instead, it is very general and solely based on the linear growth of structures in a ΛCDM cosmology.

The left panel of Fig. 3.8 shows the predictions for the projected correlation function according to our "linear theory" model, for different values of the minimum host mass M_{min}. These are compared with data in a quantitative way by determining the χ^2 statistics for each M_{min} in the left panel of Fig. 3.8. The χ^2 is computed by taking into account the covariances between different data points. We see that we obtain values of the χ^2 in the range $\chi^2 \approx 6 - 7$, which are perfectly compatible with data and translate into reduced chi-squared values of $\approx 1.5 - 1.75$. There is a slight preference in our model for smaller values of the minimum host mass, but the measurement is not statistically significant for any reasonable values of $\log_{10} M_{\mathrm{min}}/\mathrm{M}_\odot \lesssim 13.5$.

The conclusion obtained here in the context of our model differs from the one found by Arita et al. (2023), who analyzed the same data and measured a rather high value of the characteristic host halo mass for quasars at $z \approx 6$, i.e., $\log_{10} M/M_\odot = 12.9^{+0.4}_{-0.7}$. The striking difference between our conclusions and the ones in the Arita et al. (2023) analysis may reside in the different assumptions made for the shape and normalization of the correlation functions. While we assume physically-motivated halo correlation functions that follow linear theory, and convert these into a quasar-correlation function in a second step, Arita et al. (2023) parametrize the quasar auto-correlation function directly by assuming a power-law shape with a slope of -1.8 and a normalization set by the quasar auto-correlation length, $r_{0,\mathrm{QQ}}$. The results for this parametrization are also shown in Fig. 3.8 with green shadings (with the corresponding chi-squared values shown in the right panel). It is quite interesting to see that the power-law shaped models for the quasar auto-correlation functions reach a better agreement with the data than the linear theory ones, with a minimum $\chi^2 \lesssim 5$ corresponding to large values of the auto-correlation length ($r_{0,\mathrm{QQ}} \approx 20 - 50\,\mathrm{cMpc}$), in agreement with the findings of Arita et al. (2023).

We conclude by noting that our model presented in the main analysis (Sec. 3.4) is compatible with the data from Arita et al. (2023). Indeed, if we take the best-fit parameters from Fig. 3.1 and compare the prediction for the quasar auto-correlation function with data we find a value for the chi-square of $\chi^2 \approx 6$, which is consistent with the discussion above and implies a good match with observations. This implies that the Arita et al. (2023) measurements are perfectly compatible with the clustering constraints from JWST (E24). However, the Arita et al. (2023) data are very uncertain and limited only to very large scales. As a consequence, they result in rather weak constraints that – as shown in this Section – are very sensitive to the exact prescription made for the shape of the quasar auto-correlation function.

3.D Appendix: Quasar-host halo masses with a uniform luminosity threshold

As discussed in Sec. 3.5.1, the quasar host mass functions (QHMFs) shown in Fig. 3.5 are obtained by setting a luminosity threshold for modeling quasar clustering that varies with redshift according to the one employed in observations. Here, we show (Fig. 3.9) the effect of setting a uniform luminosity threshold of $\log_{10} L_{\rm thr}/{\rm erg\,s^{-1}} = 46.7$ at all redshifts. This threshold corresponds to the one employed at $z \approx 4$, so the $z \approx 4$ results are the same as in Fig. 3.5. The QHMF at $z \approx 2$ ($z \approx 6$) shifts to higher (lower) masses respectively, due to the different quasar population probed by the Eftekharzadeh et al. (2015) (E24) data. This effect, however, is not strong enough to impact in any relevant way the discussion on the evolution of quasar properties with redshift made in Sec. 3.5.1.

Acknowledgements

We are grateful to Junya Arita and the SHELLQs team for sharing their data on the quasar auto-correlation function and to Jan-Torge Schindler for discussion on the quasar luminosity function. We acknowledge helpful conversations with the ENIGMA group at UC Santa Barbara and Leiden University. EP is grateful to Rob McGibbon and Victor Forouhar Moreno for help with the simulation outputs, and to Timo Kist, Jiamu Huang, and Vikram Khaire for comments on an early version of the manuscript. JFH and EP acknowledge support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 885301). This work is partly supported by funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860744

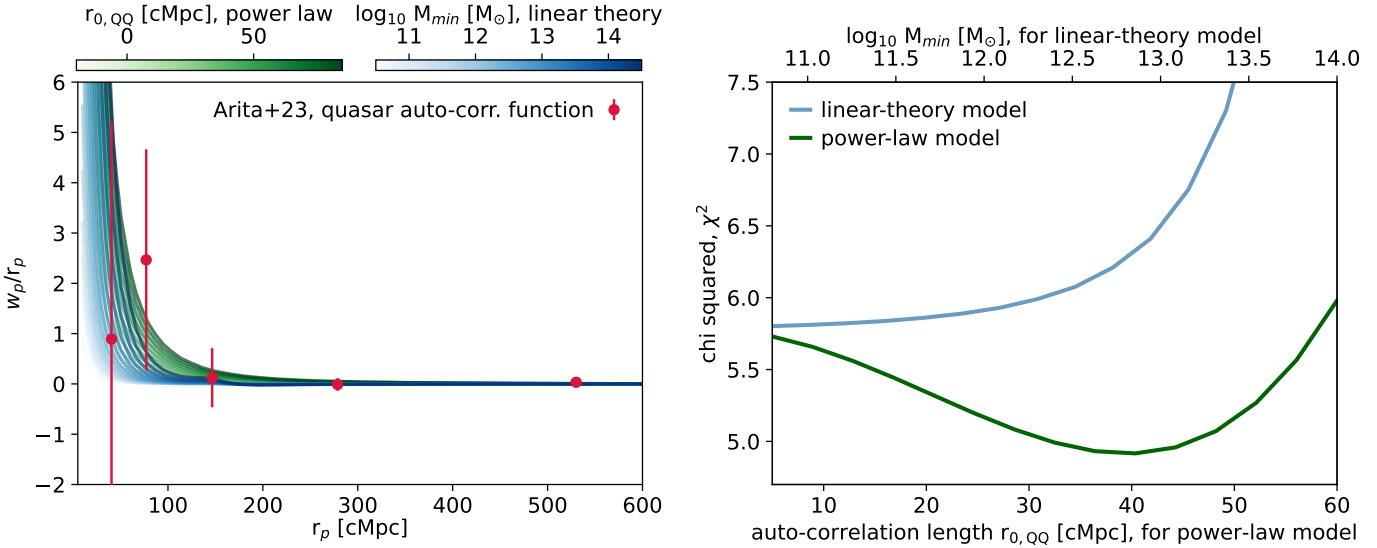

Figure 3.8: *Left*: Projected $z \approx 6$ quasar auto-correlation function, w_p/r_p as a function of the distance, r_p. The observational data from Arita et al. (2023) are shown as red points. Predictions from the model based on linear theory, which is our fiducial one, are shown as blue lines, color-coded based on the value of the $M_{\rm min}$ parameter (eq. 3.23). Predictions coming from a power-law model for the correlation functions are shown in green, color-coded according to the value of the quasar auto-correlation length, $r_{0,\rm QQ}$. *Right*: Comparison of model predictions with data, according to the value of the χ^2 statistic. The blue line refers to the "linear theory" model, and it is parametrized by the minimum host mass $M_{\rm min}$ (top label). The green line, instead, refers to the "power-law" model and is parametrized by the quasar auto-correlation length, $r_{0,\rm QQ}$ (bottom label).

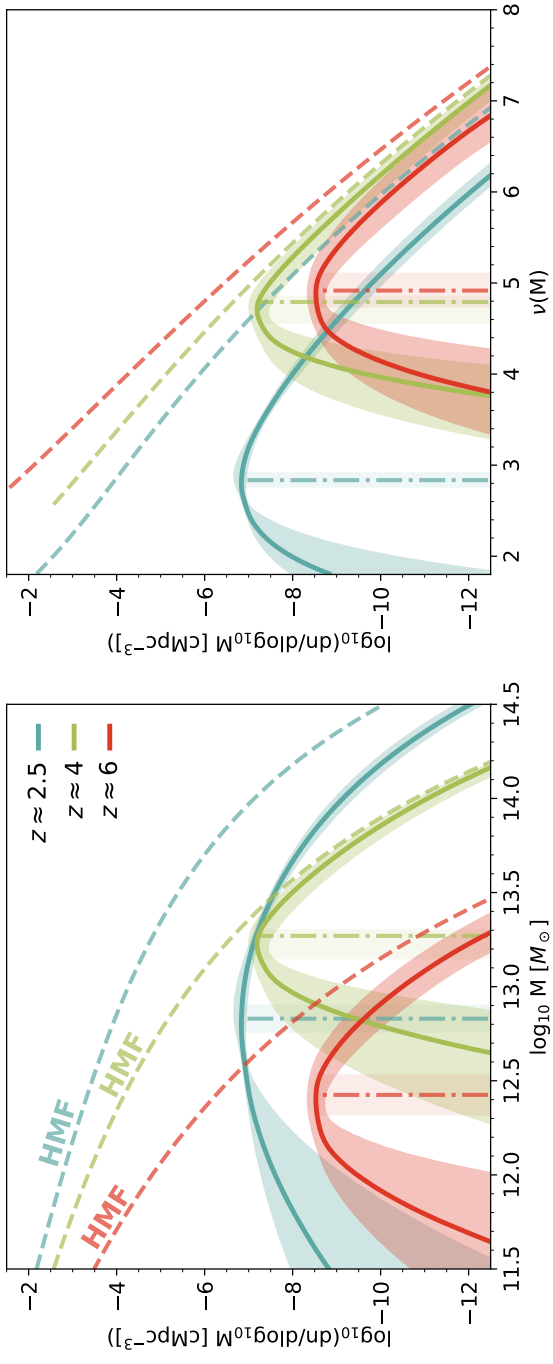

Figure 3.9: Same as Fig. 3.5, but the QHMFs here are obtained by setting a uniform luminosity threshold for the clustering measurements at all redshifts, i.e., $\log_{10} L_{\rm thr}/{\rm erg\,s^{-1}} = 46.7$. The QHMF represents the mass distribution of halos that are hosting quasars brighter than $L_{\rm thr}$.

(BiD4BESt). FW acknowledges support from NSF grant AST-2308258. This work used the DiRAC Memory Intensive service (Cosma8) at the University of Durham, which is part of the STFC DiRAC HPC Facility (www.dirac.ac.uk). Access to DiRAC resources was granted through a Director's Discretionary Time allocation in 2023/24, under the auspices of the UKRI-funded DiRAC Federation Project. The equipment was funded by BEIS capital funding via STFC capital grants ST/K00042X/1, ST/P002293/1, ST/R002371/1 and ST/S002502/1, Durham University and STFC operations grant ST/R000832/1. DiRAC is part of the National e-Infrastructure.

4 "Little red dots" cannot reside in the same dark matter halos as comparably luminous unobscured quasars

Abstract

The James Webb Space Telescope (JWST) has uncovered a new population of candidate broad-line AGN emerging in the early Universe, named "little red dots" (LRDs) because of their compactness and red colors at optical wavelengths. LRDs appear to be surprisingly abundant ($\approx 10^{-5}\,\mathrm{cMpc}^{-3}$) given that their inferred bolometric luminosities largely overlap with those of the UV-luminous quasars identified at high z in wide-field spectroscopic surveys. In this work, we investigate how the population of LRDs and/or other UV-obscured AGN relates to the one of unobscured, UV-selected quasars. By comparing their number densities, we infer an extremely large and rapidly evolving obscured:unobscured ratio, ranging from $\approx 20:1$ at $z \approx 4$ to $\approx 2300:1$ at $z \approx 7$, and possibly extending out to very high ($\approx 10^{47}\,\mathrm{erg\,s}^{-1}$) bolometric luminosities. This large obscured:unobscured ratio is incompatible with the UV-luminous duty cycle measured for unobscured quasars at $z \approx 4-6$, suggesting that LRDs are too abundant to be hosted by the same halos as unobscured quasars. This implies that either (a) the bolometric luminosities of LRDs are strongly overestimated or (b) LRDs follow different scaling relations than those of UV-selected quasars, representing a new population of accreting SMBHs emerging in the early Universe. A direct comparison between the clustering of LRDs and that of faint UV-selected quasars will ultimately confirm these findings, and shed light on key properties of LRDs such as their host mass distribution and duty cycle. We provide a mock analysis for the clustering of LRDs and show that it is feasible with current and upcoming JWST surveys.

Published in: **EP**, Joseph F Hennawi, Joop Schaye, Anna-Christina Eilers, Jiamu Huang, Jan-Torge Schindler, Feige Wang, *"Little Red Dots" cannot reside in the same dark matter halos as comparably luminous unobscured quasars*, Monthly Notices of the Royal Astronomical Society, Volume 539, Issue 4, June 2025, Pages 2910–2925, doi.org/10.1093/mnras/staf660. Reprinted here in its entirety.

4.1 Introduction

The connection between the quasar phenomenon and the accretion of material onto a supermassive black hole (SMBH) was first hypothesized to account for the extraordinary luminosity inherent to quasar activity (e.g., Salpeter 1964; Zel'dovich & Novikov 1967; Lynden-Bell 1969). According to this picture, most of the accreting material contributes to growing the mass of the SMBH, but a small fraction of this material (known as the *radiative efficiency*) is converted into energy and radiated away, giving rise to the quasar phenomenon.

The argument first proposed by Soltan (1982) embeds this connection into a cosmological context: integrating the total energy emitted by quasars over all cosmic time and assuming a standard radiative efficiency of $\approx 10\%$, one finds that the mass that has been accreted on black holes per unit of comoving volume up until today is comparable to the total mass density of the SMBHs we observe in the local Universe. This implies that SMBHs grew their mass while, at the same time, they were shining as active luminous quasars.

Extensions of this argument have been employed to relate the growth of black holes to quasar activity at different cosmic times (e.g., Yu & Tremaine 2002; Shankar et al. 2010a). While specific assumptions vary, these arguments are all based on the key idea that the bulk of black hole growth in the Universe is traced by the evolving demographic properties of luminous quasars. Wide-field optical spectroscopic surveys such as the Sloan Digital Sky Survey (SDSS, York et al. 2000) and the 2dF QSO redshift survey (2QZ, Croom et al. 2004) examined the properties of UV-luminous, type-1 quasars, and consistently showed that quasar activity peaks around $z \approx 2$ and declines rapidly towards higher redshifts (e.g., Richards et al. 2006; Kulkarni et al. 2019).

UV-luminous quasars, however, are not the whole story. The radiation emitted from accreting SMBHs can be obscured by intervening dust and gas, resulting in a diverse population of Active Galactic Nuclei (AGN) whose emission properties vary greatly across the electromagnetic spectrum (e.g., Padovani et al. 2017). A general dichotomy exists, however, between unobscured AGN/quasars, exhibiting a UV-optical continuum from the accretion disk, and obscured/reddened AGN whose UV emission is partly (or completely) extincted by the dust that surrounds the SMBH. Whether this obscuration results from a viewing-angle effect (Antonucci 1993; Urry & Padovani 1995) or signifies a distinct "dust-enshrouded" population (Sanders et al. 1988; Hopkins et al. 2005) has been hotly debated. Nevertheless, decades of AGN censuses across the electromagnetic spectrum (optical, X-ray, mid-IR, radio) have allowed us to map the contribution of UV-obscured AGN activity as a function of redshift and AGN luminosity (e.g., Ueda

et al. 2003, 2014; Merloni et al. 2014; Aird et al. 2015; Glikman et al. 2018) The resulting consensus is that a significant fraction ($\approx 20-80\%$) of AGN can be obscured in the UV, even at quasar-like (intrinsic) luminosities ($L_{\rm bol} \gtrsim 10^{45}{\rm erg\,s^{-1}}$), and that this fraction evolves mildly with redshift. Studies that include the contribution of obscured AGN environments to the total SMBH growth budget (e.g., Hopkins et al. 2007b; Shen et al. 2020) support the general picture outlined by the Soltan argument, pointing to a radiative efficiency for accretion on SMBHs close to $\approx 10\%$, and indicating that the bulk of SMBH growth took place during cosmic noon ($z \approx 1-3$).

While a multi-wavelength exploration of AGN activity is possible at $z \lesssim 3$, our understanding of black hole growth and accretion in the high-redshift Universe ($z \gtrsim 4$) has been informed almost exclusively by the population of UV-luminous, type-1 quasars detected by optical/NIR wide-field surveys up to $z \approx 7.5$ (e.g., Fan et al. 2023). This population is commonly assumed to trace the underlying evolution of AGN/SMBH activity (including UV-obscured sources) at high z by simply extrapolating the obscuration properties of quasars from low/intermediate redshifts (e.g., Shen et al. 2020). Whether this extrapolation is reliable and can offer an unbiased view of SMBH growth and AGN activity in the first billion years of the Universe is currently unclear. Several simulations (e.g., Ni et al. 2020; Vito et al. 2022; Bennett et al. 2024) and observations (Vito et al. 2018; Circosta et al. 2019; D'Amato et al. 2020; Gilli et al. 2022), for example, have suggested a rapid evolution of the obscuration properties of quasars/AGN in the early Universe, due to the presence of high-column density gas within the innermost regions of their host galaxies.

The advent of the James Webb Space Telescope (JWST) marks a huge step forward in the study of AGN activity and SMBH growth in the early Universe. JWST has the sensitivity to go beyond the UV-selected quasar population that has been studied for decades (e.g., Fan et al. 2023). Indeed, early results are already causing a seismic shift in our understanding of AGN populations at high z: photometric and spectroscopic JWST surveys are uncovering surprisingly large samples of faint AGN candidates at $z \approx 4-10$ (e.g., Harikane et al. 2023; Maiolino et al. 2024; Übler et al. 2023; Kocevski et al. 2023; Kokorev et al. 2023; Scholtz et al. 2023; Matthee et al. 2024b; Greene et al. 2024; Bogdán et al. 2024; Kocevski et al. 2024; Mazzolari et al. 2024; Furtak et al. 2024; Taylor et al. 2024). Although selection methods vary, the most reliable candidates are identified via broad Hα or Hβ lines. These lines can be used to infer AGN luminosities of $L_{\rm bol} \gtrsim 10^{44-45}\,{\rm erg\,s^{-1}}$ and black hole masses of $M_{\rm BH} \gtrsim 10^{6-7}\,{\rm M_\odot}$ These masses and luminosities vastly extend the range of AGN properties that we can probe at high z, offering key insights on the co-evolution of SMBHs and their host galaxies (e.g., Inayoshi et al. 2022; Pacucci et al. 2023), the contribution of AGN to hydrogen reionization (e.g., Maiolino et al. 2024; Dayal et al. 2024; Madau

et al. 2024), and potentially also on SMBH seeding/growth models (e.g., Pacucci & Loeb 2022; Li et al. 2024).

Yet, relating this new population of JWST AGN to the one of UV-selected high-z quasars has proven challenging. Even though they generally resemble standard, type-1 quasars at rest-frame optical wavelengths, JWST broad-line AGN appear to be much more abundant than what was expected by extrapolating the quasar luminosity function (QLF) to faint UV luminosities (Harikane et al. 2023). It is currently unclear whether QLF studies have been strongly underestimating the number of faint UV quasars that are present at high z (e.g., Giallongo et al. 2019), or whether the AGN population revealed by JWST using broad optical lines presents substantially different properties from those of UV-selected, type-1 quasars, as also suggested by their peculiar Spectral Energy Distribution (SED) features such as X-ray weakness (Maiolino et al. 2024; Lambrides et al. 2024a) and (tentative) lack of variability (Kokubo & Harikane 2024). Upcoming JWST surveys will probe the properties of these broad-line AGN in the rest-frame UV, providing key insight into their nature and allowing a direct comparison to the UV-selected quasar population.

Interestingly, however, some of the AGN revealed by JWST are even more remarkable: a significant fraction of them ($\gtrsim 20\%$; Harikane et al. 2023; Taylor et al. 2024) show a steep red continuum in the rest-frame optical pointing to moderate dust reddening values of $A_V \approx 1-4$ (Kokorev et al. 2024a; Greene et al. 2024). When correcting for the attenuation of dust to the continuum and/or broad-line emission, these obscured/reddened AGN have inferred bolometric luminosities of $L_{\rm bol} \approx 10^{45-46}\,{\rm erg\,s^{-1}}$ and SMBH masses up to $\approx 10^{7-8}\,{\rm M_\odot}$ (Greene et al. 2024; Kocevski et al. 2024; Harikane et al. 2023). Hence, they largely overlap in luminosity and SMBH mass with the population of UV-selected, type-1 quasars revealed in pre-JWST surveys (Fan et al. 2023; Matsuoka et al. 2022). This is incredibly surprising, since these UV-luminous quasars with comparable luminosities (and redshifts) were selected from wide-field 1400 deg^2 deep imaging surveys probing volumes of $\approx 10^{10}$ cMpc^3 (Matsuoka et al. 2022), whereas JWST AGN are identified in surveys of not more than $\approx 300-600$ arcmin^2 probing a volume not greater than $\approx 10^6-10^7$ cMpc^3 (Matthee et al. 2024b; Kokorev et al. 2024a). Such a massive difference indicates that these AGN may be tracing a new population of broad-line, obscured sources[1] that are far more abundant than

[1]Standard AGN classifications (e.g., Padovani et al. 2017) divide low-z quasars in type-1 (showing broad emission lines in their spectra) and type-2 (showing only narrow emission lines). Type-2 quasars are generally identified with obscured sources whose broad lines are extincted by dust. Even though their continuum is heavily reddened at optical and UV wavelengths, JWST AGN are always revealed by broad optical lines, and hence they officially belong to the type-1 quasar category. While examples of type-1, reddened quasars exist at low redshifts, they are rare compared to the global quasar population (Wang et al., in prep.), making the interpretation of these new JWST AGN sources even more challenging.

comparably luminous UV-unobscured quasars. According to this picture, our understanding of SMBH growth and quasar/AGN activity at high-z – which was entirely based on the demographic properties of UV-luminous quasars – needs to be thoroughly revised to account for this new, large AGN population that is in place in the early Universe (e.g., Inayoshi & Ichikawa 2024; Li et al. 2025a).

As shown by Greene et al. (2024), the reddened broad-line AGN in JWST surveys tend to have a characteristic v-shaped SED, with the red continuum in the rest-frame optical transitioning to relatively blue colors in the rest-frame UV. While the physical origin of this SED shape is currently unclear (e.g., Killi et al. 2024; Li et al. 2025a; Wang et al. 2024; Kokorev et al. 2024b; Inayoshi & Maiolino 2025), several studies have exploited these peculiar SED features and applied specific color and compactness cuts to NIRCam photometry to isolate obscured broad-line AGN photometrically (e.g., Labbe et al. 2025; Pérez-González et al. 2024; Kokorev et al. 2024a; Kocevski et al. 2024; Akins et al. 2024). By applying similar photometric selections, Greene et al. (2024) and Kocevski et al. (2024) have proved that a large fraction of the selected sources ($\gtrsim 70-80\%$) is indeed comprised of reddened, high-redshift ($z \approx 4-8$), broad-line AGN. Sources selected using these methods have become known as "Little Red Dots" (LRDs henceforth; Matthee et al. 2024b) because of their compactness and peculiar colors in NIRCam imaging. We note that this term has been used in the literature to refer to samples obtained following different spectroscopic and photometric criteria. Here, with the term "Little Red Dots" we refer to the above-mentioned population of candidate broad-line AGN that are red at optical wavelengths, and hence have quasar-like inferred bolometric luminosities and black hole masses. We include in our analysis both spectroscopic (Greene et al. 2024) and photometric (Kokorev et al. 2024a) samples: while the latter may be subject to a significant degree of contamination (e.g., Taylor et al. 2024), their number densities agree well with the ones from spectroscopy (Greene et al. 2024)[2]. We mention the caveat, however, that even for spectroscopically confirmed broad-line LRDs, the presence of an accreting SMBH and the nature of the observed SED are still heavily debated (e.g., Durodola et al. 2024; Li et al. 2025a; Pérez-González et al. 2024; Ananna et al. 2024; Yue et al. 2024b; Maiolino et al. 2024; Kokubo & Harikane 2024; Baggen et al. 2024; Inayoshi & Maiolino 2025). In the following, we *assume* that LRDs are obscured, broad-line AGN, and examine the consequences of the large obscured:unobscured ratio at quasar-like bolometric luminosities that is implied by this assumption. We defer the reader to Sec. 4.5 for a discussion on the nature of LRDs and the conclusions we can draw from our results. There, we will also examine how the general population of faint (unobscured)

[2]On top of that, a moderate degree of contamination does not impact the main conclusions of our analysis (see Sec. 4.5 for further discussion).

broad-line AGN revealed by JWST (e.g., Harikane et al. 2023; Maiolino et al. 2024; Taylor et al. 2024) fits in the discussion presented in this work.

If a huge obscured LRD population is indeed present at high redshifts, the first question that awaits to be answered is: how does this population compare to that of comparably luminous, UV-selected quasars in terms of SMBH mass and accretion rate, host environments, and evolution history? Are LRDs standard, actively accreting quasars whose emission is attenuated by intervening dust and gas, or do they represent a different evolutionary stage in the accretion history of SMBHs? Are UV-luminous quasars and LRDs drawn from the same population of halos/galaxies?

In this work, we take a first step towards answering these questions by studying the properties of quasars and LRDs in terms of their number density and large-scale environment/host halo mass. In particular, we argue that the extreme abundance of LRDs/obscured AGN is at odds with the duty cycle of UV-luminous quasar activity at $z \approx 4-6$ inferred from the combination of quasar clustering and luminosity function measurements (Shen et al. 2007; Eilers et al. 2024; Pizzati et al. 2024a,b). This indicates that LRDs cannot be drawn from the same population of dark matter halos as UV-selected quasars, notwithstanding that quasars and LRDs have the same inferred bolometric luminosities and SMBH masses. Hence, provided that these luminosities and masses are indeed correct, LRDs would need to obey fundamentally different scaling relations than the ones holding for quasars, as the same SMBH masses are linked to smaller host halo masses. Possibly, this points to the fact that LRDs represent a different evolutionary stage in the accretion history of SMBHs at early cosmic time.

In order to support these conclusions and unveil the accretion history and large-scale environment of LRDs, measuring the clustering of these sources is key. Here, we suggest that a convincing measurement of the duty cycle and host halo mass of LRDs can be obtained by using NIRCam/WFSS observations of LRD fields and measuring the cross-correlation between LRDs and [O III] line emitters, with a similar setup and strategy to current JWST programs targeting UV-luminous, high-z quasars, such as EIGER (Kashino et al. 2023; Eilers et al. 2024) and ASPIRE (Wang et al. 2023). Using the methodology developed in previous work (Pizzati et al. 2024a,b, Chapters 2-3), we provide a mock analysis for these clustering measurements and discuss the prospect of undertaking this measurement with current and future JWST programs.

The paper is structured as follows. In Sec. 4.2, we compare the abundance of LRDs/obscured AGN with the one of the UV-luminous high-z quasar population, inferring a large and rapidly evolving obscured:unobscured ratio at $z \approx 4-8$. Sec. 4.3 studies the implications of this large ratio in terms of host dark matter halo populations, and points to clustering studies as a way to determine the nature of LRDs. Sec. 4.4 provides a mock analysis of this

clustering measurement. The results are discussed and summarized in Sec. 4.5.

4.2 The staggeringly high abundance of UV-obscured AGN implied by little red dots

In this section, we compare the luminosity function of the UV-luminous, unobscured population of quasars to that of the new population of UV-obscured "Little Red Dots" (LRDs) uncovered in JWST surveys. Our goal is to study the abundance of these two populations across cosmic time, and infer an estimate of the AGN obscured fraction at different redshifts.

To this end, we use bolometric luminosities as a way to probe the intrinsic radiation emitted by the different quasar/AGN populations prior to any obscuration effects. The bolometric luminosities of UV-luminous, type-1 quasars can be easily inferred from their UV-continuum absolute magnitude by assuming standard bolometric correction factors that are available in the literature (e.g., Richards et al. 2006; Runnoe et al. 2012a; Shen et al. 2020). In this work, we use the relation between the M_{1450} absolute magnitude and the bolometric luminosity $L_{\rm bol}$ presented in Runnoe et al. (2012a)[3]. While other bolometric correction factors may return slightly different results because of the choices made for the quasar SED and the parametrization of the UV-bolometric relation, the uncertainty in the bolometric correction for UV-selected, type-1 quasars is relatively small and has little impact on our conclusions.

Estimating the intrinsic bolometric luminosity of the LRD population, instead, is much more challenging. While bolometric luminosities are easy to constrain for UV-selected quasars because one directly probes the big-blue-bump (where the bulk of the emission comes out, Sanders et al. 1989), dust obscuration prevents a direct determination of the LRD luminosities from their UV emission. For low-z, dust-obscured quasars, it is usually possible to constrain the radiation reprocessed by dust in the mid-IR with Spitzer (e.g., Lacy et al. 2015). However, this is currently not a viable option for LRDs, as they appear to manifest only at high z and the bulk of their expected mid-IR emission is redshifted to wavelengths of $\approx 70\mu$m, which are not accessible from the ground and are only probed by shallow surveys (e.g., Herschel). The only option that remains available for estimating the bolometric luminosities of LRDs is to use the emission in the optical continuum and/or broad optical lines and convert that to a bolometric luminosity using some scaling relations (e.g., Richards et al. 2006; Runnoe et al. 2012b), which are however

[3]The bolometric correction for $\lambda = 1450$ Å is $\log_{10} L_{\rm iso}/{\rm erg\,s^{-1}} = 4.745 + 0.910 \log_{10} \lambda L_\lambda/{\rm erg\,s^{-1}}$. $L_{\rm iso}$ refers to the bolometric luminosity computed under the assumption of isotropy, and it is related to the observed bolometric luminosity $L_{\rm bol}$ through the relation $L = 0.75\, L_{\rm iso}$.

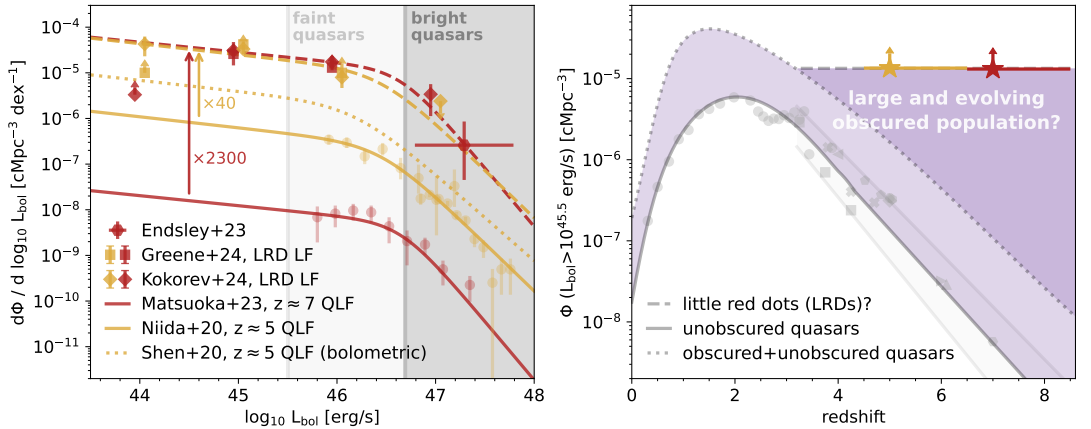

Figure 4.1: *Left*: Luminosity function (LF) of UV-selected quasars, expressed in terms of bolometric luminosities, compared to the bolometric LF of Little Red Dots (LRDs) at different redshifts. Solid lines show the fits to the unobscured quasar luminosity functions (QLFs) at $z \approx 5$ (Niida et al. 2020; golden color) and $z \approx 7$ (Matsuoka et al. 2023; red). Data points for these QLFs are also shown as circles. The bolometric QLF compiled by Shen et al. (2020) at $z \approx 5$ is shown with a dotted line. Bolometric LFs for LRDs are shown with square (Greene et al. 2024) and diamond (Kokorev et al. 2024a) symbols. Golden (red) symbols refer in this case to the redshift range $4.5 < z < 6.5$ ($6.5 < z < 8.5$). The number density implied by the single source identified by Endsley et al. (2022, 2023) at $z \approx 7$ (see main text) is shown as a red hexagon. Vertical arrows show by how much the QLF fits (solid lines) need to be rescaled to match the LRD LF. Dashed lines show the rescaled QLFs: the $z \approx 5$ ($z \approx 7$) QLF is rescaled by a factor of 40 (2300). The light (dark) grey shaded region highlights the luminosity range $10^{45.5}\,\mathrm{erg\,s^{-1}} < L_{\rm bol} < 10^{46.5}\,\mathrm{erg\,s^{-1}}$ ($L_{\rm bol} > 10^{46.5}\,\mathrm{erg\,s^{-1}}$). *Right:* Evolution of the number density of quasar/AGN (above the luminosity threshold $L_{\rm bol} > 10^{45.5}\,\mathrm{erg\,s^{-1}}$) with redshift. Gray points show the number densities obtained by integrating individual fits to the unobscured QLFs above the luminosity threshold (partly adapted from the compilation in Schindler et al. 2023). The solid line shows an evolutionary model for the unobscured quasar number density obtained by smoothly interpolating between the fit of Kulkarni et al. (2019) at $z < 4$, and an exponential decline $\Phi \propto 10^{-kz}$, with $k = 0.7$, at higher redshifts. The gray shaded area and the two lines at $z > 3$ are meant to bracket our uncertainty on the number density of high-z unobscured quasars. Dotted lines show the number density evolution predicted by the bolometric QLF of Shen et al. (2020) (see their "global fit B"). Colored star symbols show the number density for the LRDs obtained by integrating the rescaled QLFs from the left panel. The flat evolution of the LRD number density implied by the data points is highlighted with a horizontal dashed line, while the (light+dark) purple-shaded areas show the AGN obscured:unobscured ratio inferred from LRDs and low-z multi-wavelength observations.

fairly uncertain. Even more relevantly, one has to properly account for the effects of dust obscuration on the observed optical emission. Current estimates of the bolometric luminosities for the LRD population (e.g., Greene et al. 2024; Kokorev et al. 2024a; Akins et al. 2024) rely on the assumption that the optical continuum of LRDs is dominated by dust-reddened AGN radiation and use the slope of the SED in the optical continuum to infer the amount of obscuration in place. However, this continuum emission could be contaminated by radiation from the host galaxy: disentangling the contributions of the central SMBH and the stellar light to the SED of LRDs is currently a hotly debated problem (e.g., Durodola et al. 2024; Li et al. 2025a; Pérez-González et al. 2024; Baggen et al. 2024; Inayoshi & Maiolino 2025). As mentioned before, here we simply assume that bolometric luminosity estimates for LRDs are correct. A discussion on how our results are impacted by uncertainties in the bolometric luminosities of LRDs can be found in Sec. 4.5.

In the left panel of Figure 4.1, we show the luminosity function of UV-luminous, unobscured quasars (expressed in terms of bolometric luminosities) at two sample redshifts of $z \approx 5$ (Niida et al. 2020; golden solid line and points) and $z \approx 7$ (Matsuoka et al. 2023; red solid line and points). These luminosity functions can be compared to the bolometric luminosity functions of LRDs measured by Greene et al. (2024) (squares) and Kokorev et al. (2024a) (diamonds)[4]. Golden (red) symbols refer to the redshift range $4.5 < z < 6.5$ ($6.5 < z < 8.5$). This plot highlights the strikingly different abundance of LRDs compared to the UV-luminous quasar population. As also mentioned in the introduction, this difference reflects the fact that LRDs are common in the small fields ($\approx 300 - 600$ arcmin^2) probed by JWST surveys, whereas unobscured quasars are notoriously rare and can be sampled only by wide-field surveys of ≈ 2000 deg^2.

By directly comparing the luminosity functions of UV-luminous quasars and LRDs, we can quantify the different abundances of these two populations as a function of their luminosity. Interestingly, we find that the *shape* of

[4]The Greene et al. (2024) luminosity function is obtained from a small sample of spectroscopically-confirmed broad-line LRDs in the UNCOVER field (Bezanson et al. 2024). The work of Kokorev et al. (2024a) applies the photometric selection suggested by Labbe et al. (2025) and Greene et al. (2024) to a larger sample of JWST blank fields, identifying 260 AGN candidates in ≈ 640 arcmin^2 of JWST imaging. While several other LRD luminosity functions have been published in the literature (see e.g., Matthee et al. 2024b; Kocevski et al. 2024; Lin et al. 2024), none of these are based on *unattenuated* bolometric luminosities. Accounting for the effect of dust attenuation is key if our goal is to compare the luminosities of LRDs to the ones of UV-luminous quasars. The only exception is the recent work of Akins et al. (2024), who also published an LRD bolometric luminosity function corrected for obscuration effects. However, their photometric selection differs significantly from the one presented in Greene et al. (2024) and Kokorev et al. (2024a), and hence we do not include their sample in the analysis. We note however that they find even larger number densities for LRDs, which would strengthen our conclusion on the presence of a large obscured high-z AGN population.

the LRD luminosity function resembles the one of the UV-luminous quasar luminosity function (QLF) at both redshifts. Indeed, if we scale up the Niida et al. (2020) fit to the $z \approx 5$ QLF by a factor of ≈ 40, we get a good match to the LRD luminosity function in the redshift range $4.5 < z < 6.5$. This suggests that LRDs may constitute a new, obscured population of accreting SMBHs at $z \approx 5$, outnumbering unobscured quasars by ≈40:1 at all luminosities. Similar – but even more extreme – conclusions can be drawn at $z \approx 7$. In this case, the fit to the Matsuoka et al. (2023) QLF needs to be scaled up by a factor of ≈ 2300 to match the LRD luminosity function at $6.5 < z < 8.5$, implying an even larger obscured:unobscured ratio, roughly independent of luminosity.

We note that care must be taken to extend these conclusions to a large range of bolometric luminosities. Most LRDs have inferred (dust-corrected) bolometric luminosities in the range $\approx 10^{44-46}\,\mathrm{erg\,s^{-1}}$. The faintest high-$z$ unobscured quasars identified in wide field surveys have luminosities of $\approx 10^{45.3}\,\mathrm{erg\,s^{-1}}$ (e.g., Matsuoka et al. 2022). Hence, a proper comparison between LRD and quasar number densities can be carried out only for the *bright* population of LRDs with $L_{\mathrm{bol}} \approx 10^{45.5-46.5}\,\mathrm{erg\,s^{-1}}$. At lower bolometric luminosities, the UV-luminous QLFs are only based on extrapolations; hence, conclusions on the obscured fraction of faint ($L_{\mathrm{bol}} \lesssim 10^{45}\,\mathrm{erg\,s^{-1}}$) AGN are only tentative. At very bright luminosities of $L_{\mathrm{bol}} \approx 10^{47}\,\mathrm{erg\,s^{-1}}$, the number density of UV-luminous quasars is very well constrained (e.g., Schindler et al. 2023). Very bright LRDs, on the other hand, are hard to find in the small field of views (FoVs) probed by JWST surveys and the only constraints we have on their number density come from the work of Kokorev et al. (2024a) (see also Akins et al. 2024), which is however only based on photometry with no spectroscopic confirmation.

Interestingly, signs of a large obscured AGN population at high bolometric luminosities ($L_{\mathrm{bol}} \gtrsim 10^{47}\,\mathrm{erg\,s^{-1}}$) come from different data. Using multi-wavelength observations in mid-/far-IR, sub-mm, and radio, Endsley et al. (2022, 2023) (see also Lambrides et al. 2024b) discovered an extremely luminous ($L_{\mathrm{bol}} = (2.0 \pm 0.2) \times 10^{47}\,\mathrm{erg\,s^{-1}}$) obscured, radio-loud quasar at $z = 6.83$ in just $1.5\,\mathrm{deg}^2$ of COSMOS imaging, and argued for an extremely large obscured:unobscured ratio of $\sim 2000 : 1$. We can get an estimate of the number density implied by this source by simply computing the total comoving volume in the COSMOS field for the redshift range $6.6 < z < 6.9$ (in which the source was photometrically selected; see Endsley et al. 2022). We get a volume of $3.8 \times 10^6\,\mathrm{cMpc}^3$ and a number density of $2.6 \times 10^{-7}\,\mathrm{cMpc}^{-3}$. For reference, we add this source to the luminosity function plot of Fig. 4.1 (left), by assuming a 1 dex bin in bolometric luminosity centered on the quasar's measured L_{bol}. Upper and lower limits are computed assuming Poisson statistics for a single source (see Gehrels 1986). Despite the large uncertainties, this source supports the existence of

a large obscured population at the bright end of the QLF compatible with the one found for LRDs.

In what follows, we will consider two separate hypotheses: (a) there is a large obscured AGN/quasar population at bolometric luminosities $L_{\rm bol} \approx 10^{45} - 10^{46}\,{\rm erg\,s^{-1}}$ (i.e., at the faint end of the quasar luminosity function; light-grey shaded area in the left panel of Fig. 4.1); (b) this large obscured population extends to very large bolometric luminosities of $L_{\rm bol} \approx 10^{47}\,{\rm erg\,s^{-1}}$ (dark-grey shaded area). While the former is supported by a fairly large sample of LRDs that have been argued to overlap in luminosity with the faint quasar population (e.g., Greene et al. 2024; Matthee et al. 2024b; Lin et al. 2024; Taylor et al. 2024; Schindler et al, in prep.), the latter is currently based only on a handful of sources (i.e., the photometrically-selected LRDs in Kokorev et al. 2024a; Akins et al. 2024 and the obscured quasars from Endsley et al. 2022, 2023; Lambrides et al. 2024b) and thus it is only tentative (see Sec. 4.5 for further discussion).

In the right panel of Figure 4.1, we show how the quasar/AGN number density evolves with redshift by integrating the QLF above a bolometric luminosity threshold of $L_{\rm bol} = 10^{45.5}\,{\rm erg\,s^{-1}}$ (light grey vertical line in the left panel). The cosmological evolution of the UV-luminous, type-1 quasar population has been analyzed in the recent work of Kulkarni et al. (2019). The solid grey line in Fig. 4.1 (right) shows their best-fitting model at $z < 4$. For higher redshifts, the Kulkarni et al. (2019) model is very uncertain and does not agree well with the data. For this reason, at $z > 4$ we assume that the cosmic number density of unobscured high-z quasars declines exponentially as $\Phi(z) \propto 10^{-kz}$, and set $k = 0.7$ for our fiducial model (Schindler et al. 2023). We then smoothly interpolate between the fit of Kulkarni et al. (2019) at $z < 4$ and this exponential decrease at higher redshift. Together with this global evolution model, we also show individual (gray) points obtained by integrating local fits to the QLFs above the luminosity threshold (fits are taken from Yang et al. 2016; Akiyama et al. 2018; McGreer et al. 2018; Matsuoka et al. 2018; Schindler et al. 2019; Kulkarni et al. 2019; Niida et al. 2020; Onken et al. 2022; Pan et al. 2022; Schindler et al. 2023; Matsuoka et al. 2023). Overall, these individual data points agree with the global evolutionary model, but a significant spread is present due to uncertainties in the QLF measurements (especially at the faint end, $L_{\rm bol} \lesssim 10^{46}\,{\rm erg\,s^{-1}}$). To quantify this uncertainty, we plot two gray lines corresponding to different exponential declines of the quasar number density, $k = 0.65$ and $k = 0.78$ (e.g., Wang et al. 2019; Matsuoka et al. 2023); these two lines are normalized at $z = 4$ to twice and half of the fiducial model, respectively.

Together with the measurements for the UV-luminous quasar number density, we show (Fig. 4.1, right panel) with a dotted line the model for the evolution of the AGN *bolometric* number density from Shen et al. (2020). This work employs multi-wavelength observations (from X-rays to mid-IR) to

include the contribution of all quasars/AGN to the number density budget. In particular, by exploiting X-ray observations at $0 < z < 3$ (e.g., Ueda et al. 2003, 2014; Merloni et al. 2014; Aird et al. 2015), they include a model for AGN obscuration, and account for the obscured fraction of quasars/AGN in their luminosity function estimates. As mentioned in the introduction, observations generally constrain the AGN obscured fraction only at $z \lesssim 3$, so the Shen et al. (2020) model is effectively extrapolating the behaviour of the AGN obscured populations from cosmic noon to the high z Universe. Nonetheless, the work of Shen et al. (2020) represents our best guess (prior to JWST observations) for how the global AGN/SMBH population evolves as a function of redshift. By comparing the number density of UV-selected quasars (solid grey line in the right panel of Fig. 4.1) with the number density from Shen et al. (2020) (which includes obscured sources), we can estimate the obscured:unobscured ratio of AGN as a function of redshift. The same ratio can be studied as a function of intrinsic luminosity by considering the UV-luminous and the bolometric QLFs at a single redshift. As an example, we do this in the left panel of Fig. 4.1 by showing the Shen et al. (2020) predictions for the bolometric QLF at $z = 5$ with a golden dotted line. In general, the obscured:unobscured ratio implied by comparing the bolometric (Shen et al. 2020) to the UV (Kulkarni et al. 2019) QLFs evolves moderately with redshift and luminosity, ranging from $\approx$ a few: 1 up to $\approx 20:1$ for the case of high redshift and low bolometric luminosity. We note that these values are inevitably very uncertain, as the method employed here is subject to the exact parametrizations employed by Kulkarni et al. (2019) and Shen et al. (2020) for their respective QLFs. Nevertheless, we present this comparison between UV-selected and bolometric models to outline the conventional wisdom on AGN/quasar populations that is being challenged by the new population of LRDs/broad line AGN uncovered in JWST surveys.

The number density evolution of LRDs can be estimated by integrating their bolometric luminosity functions in the left panel of Fig. 4.1 above the same $L_{\rm bol}$ threshold of $10^{45.5}\,{\rm erg\,s^{-1}}$ employed before (vertical light grey line). In practice, given that the rescaled UV QLFs (dashed lines in the left panel of Fig. 4.1) are good fits to the LRD bolometric luminosity functions, we can simply rescale the unobscured quasar number density obtained at $z = 5$ and $z = 7$ to get the LRD number densities at the same redshifts. We show as colored star symbols (Fig. 4.1, right panel) the LRD number densities obtained after this rescaling. Following Greene et al. (2024), we plot these symbols as lower limits.

As argued before, the AGN number density implied by JWST observations of LRDs is surprisingly large and non-evolving. To highlight this behavior, we plot (Fig. 4.1, right panel) a horizontal dashed line for $z \gtrsim 3$ corresponding to the abundance $\Phi_{\rm LRD} \approx 1.3 \times 10^{-5}\,{\rm cMpc^{-3}}$. At $z \gtrsim 6$, this abundance is many orders of magnitude higher than the one measured for unobscured quasars, implying that our general understanding of SMBH

accretion and quasar activity in the early Universe may need to be deeply revised. Inayoshi & Ichikawa (2024) (see also Akins et al. 2024) have already examined the challenges that these LRD number densities pose to our paradigm of SMBH growth as well as the co-evolution of SMBHs and galaxies. In this work, we focus on the consequences of the large and rapidly evolving AGN obscured fraction that can be inferred by comparing LRDs to unobscured quasars. In Fig. 4.1 (right), we show with a light purple shading the region between the unobscured quasar evolution model and the bolometric (obscured+unobscured) model of Shen et al. (2020). A darker shading highlights the dramatic increase in the obscured fraction at $z \gtrsim 4$ that is needed to match LRD measurements.

Dividing the LRD number density, $\Phi_{\rm LRD}$ (which, to a first approximation, is not evolving with redshift), by the number density of UV-luminous quasars (solid grey line in the left panel of Fig. 4.1), we infer an obscured:unobscured ratio that increases from $r_{\rm obsc} \approx 20^{+20}_{-10} : 1$ at $z = 4$ to $r_{\rm obsc} \approx 2300^{+3500}_{-1400} : 1$ at $z = 7$. In the following section, we will also make use of the obscured:unobscured ratio at $z = 6.25$, which is $r_{\rm obsc} \approx 815^{+1600}_{-545} : 1$. The uncertainties on these obscured ratios are computed by considering the grey shaded area (and grey lines) in Fig. 4.1 (right), and are meant to quantify the scatter (coming from systematics in the QLF modeling) between different number density measurements for the unobscured quasar population. Given the challenges with interpreting and contextualizing LRD measurements, we currently do not attempt to model uncertainties for the LRD population, and defer to Sec. 4.5 for a discussion of the significance of our results.

4.3 Little red dots and UV-selected quasars: do they belong to the same population?

From the analysis performed in the previous section, we concluded that: (a) LRDs imply the existence of a large and rapidly evolving obscured AGN population (at redshifts $z \approx 4-7$ and bolometric luminosities $L_{\rm bol} \approx 10^{45}-10^{46}\,{\rm erg\,s^{-1}}$) which outnumbers UV-luminous quasars by several orders of magnitude (Fig. 4.1, right); (b) there is tentative evidence (Fig. 4.1, left) that this obscured population extends to even higher bolometric luminosities ($L_{\rm bol} \approx 10^{47}\,{\rm erg\,s^{-1}}$). In this section, we examine the implications of these findings in the context of AGN host dark matter halo masses and duty cycles.

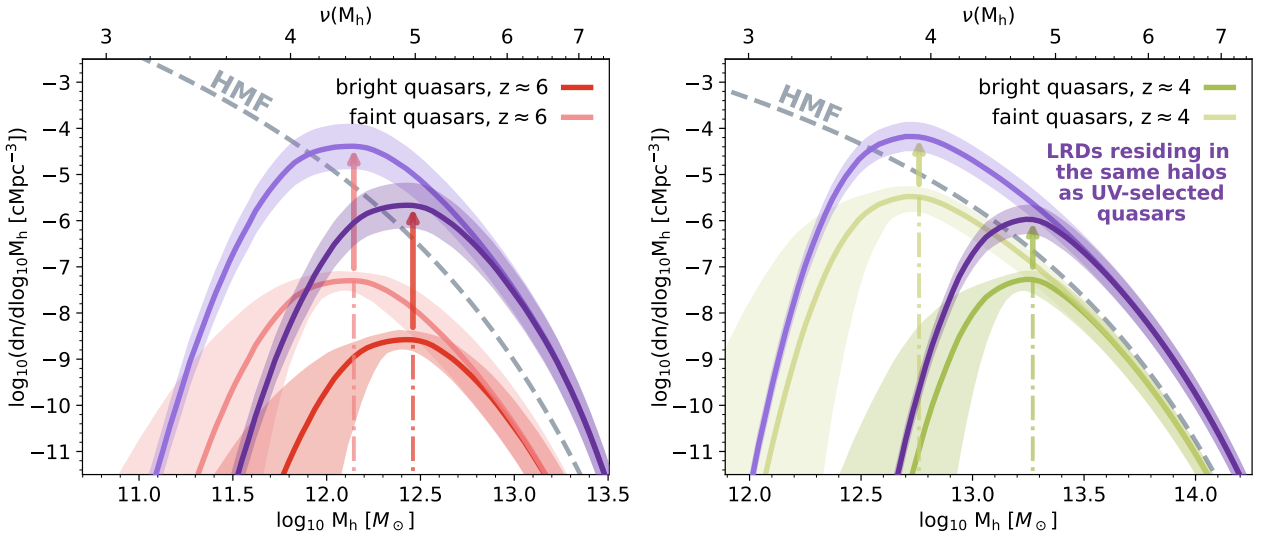

Figure 4.2: Quasar-host mass functions (QHMFs) at $z \approx 6$ (*left panel*) and $z \approx 4$ (*right*) as a function of the (sub)halo mass, $M_{\rm h}$, and the peak height, $\nu(M_{\rm h})$ (see text for definitions). Darker (lighter) colors represent QMHFs obtained by setting a bolometric luminosity threshold corresponding to that of bright (faint) quasars, i.e., $L_{\rm bol} = 10^{46.7}\,{\rm erg\,s^{-1}}$ ($L_{\rm bol} = 10^{45.5}\,{\rm erg\,s^{-1}}$). Median and 1σ uncertainties (obtained by randomly sampling the posterior distributions shown in Pizzati et al. 2024a,b), are represented with solid lines and shaded areas, respectively. The dashed-dotted lines highlight the median values of the QHMF distributions. The (sub)halo mass functions (HMFs) at the respective redshifts are plotted with dashed grey lines in both panels. Purple colors show the QHMFs distributions when scaled up by the obscured:unobscured ratios ($r_{\rm obsc}$) derived in Sec. 4.2, and represent the host mass distribution of LRDs under the hypothesis that they are drawn from the same halo population as UV-selected quasars. Dark (light) purple is associated with bright (faint) quasar bolometric luminosities. The colored arrows represent the values of $r_{\rm obsc}$ by which the QHMFs are scaled up. The (purple) shaded regions represent the effect of the uncertainties on $r_{\rm obsc}$ (see Sec. 4.2). The host mass distributions for LRDs overshoot the HMFs at the massive end, implying that LRDs are too abundant to reside in the same dark matter halos as comparably luminous, unobscured quasars.

4.3.1 The host dark matter halos and duty cycles of high-z unobscured quasars and their luminosity dependence

Determining which halos can host quasar activity as a function of cosmic time is one of the main questions in the field, as it is key to embedding quasars in the structure formation picture: this sheds light on the processes governing SMBH accretion and growth, as well as the co-evolution between SMBHs and their host halos/galaxies. In this context, quasar clustering measurements have been widely used to estimate the masses of the halos hosting UV-luminous quasars at different redshifts (Porciani et al. 2004; Croom et al. 2005; Porciani & Norberg 2006; Shen et al. 2007; Ross et al. 2009; Eftekharzadeh et al. 2015; Arita et al. 2023; Eilers et al. 2024). The idea behind these measurements is straightforward: according to the ΛCDM cosmology, the clustering of any populations of objects increases with the masses of the dark matter halos they reside in (e.g. Kaiser 1984; Bardeen et al. 1986; Mo & White 1996).

As pointed out by, e.g., Martini & Weinberg (2001); Haiman & Hui (2001), determining the quasars' characteristic host halo masses can also give us insight into their accretion history. Suppose that – as routinely assumed – all massive halos host a SMBH at their center. The *duty cycle* of quasar activity determines what fraction of these SMBHs, on average, are active as UV-luminous quasars at any given moment. By comparing the number density of potential quasar hosts – obtained from quasar clustering measurements – to the observed unobscured quasar number density, one can constrain this *UV-luminous quasar duty cycle*. Given the connection between quasar activity and SMBH accretion and growth, the quasar duty cycle offers a direct view into the growth mode of SMBHs at a given cosmic epoch.

In Pizzati et al. (2024a,b), we developed a method to constrain the UV-luminous quasar duty cycle ($\varepsilon_{\rm QSO}$) as well as the mass distribution of the (sub)halos that host unobscured quasars (the so-called "*quasar host mass function*"; QHMF) by simultaneously fitting the clustering of quasars and their luminosity function. The method builds on a conditional luminosity function (CLF) framework, which links in a statistical sense the population of dark matter subhalos to that of quasars (e.g., Yang et al. 2003; Ren et al. 2020). We employ a description for the CLF based on an empirical relation between the quasar bolometric luminosity, $L_{\rm bol}$, and the host (sub)halo mass, $M_{\rm h}$, with log-normal scatter, σ. This relation is also normalized by an active

fraction, $f_{\rm on,UV}$, which accounts for the fact that not all quasars are actively accreting and UV-luminous at a given time:

$$\begin{aligned}\mathrm{CLF}(L_{\rm bol}|M_{\rm h})\,\mathrm{d}L_{\rm bol} &= \\ &= \frac{f_{\rm on,UV}}{\sqrt{2\pi}\sigma}\,\exp\left(\frac{(\log_{10}L_{\rm bol}-\log_{10}L_{\rm c}(M_{\rm h}))^2}{2\sigma^2}\right)\mathrm{d}\log_{10}L_{\rm bol}.\end{aligned} \tag{4.1}$$

We assume a power-law $L_{\rm c}(M_{\rm h})$ relation, parametrized by a slope, γ, and a normalization $L_{\rm ref}$. In terms of logarithmic quantities:

$$\log_{10}L_{\rm c}(M_{\rm h}) = \log_{10}L_{\rm ref} + \gamma\,(\log_{10}M_{\rm h} - \log_{10}M_{\rm ref})\,, \tag{4.2}$$

with $M_{\rm ref}$ fixed to $\log_{10}M_{\rm ref}/\mathrm{M}_\odot = 12.5$.

By fitting the quasar clustering and the QLF at any given redshift, we have enough information to constrain the quasar luminosity-halo mass relation (γ and $L_{\rm ref}$), its intrinsic scatter (σ), and the active fraction of quasars ($f_{\rm on,UV}$) – see Table 4.1. Once these quantities are known, the QHMF can be obtained by statistically assigning quasars to subhalos and selecting only the subhalos whose quasars are brighter than some luminosity threshold, $L_{\rm thr}$ (which is usually set according to observations):

$$\mathrm{QHMF}(M_{\rm h}|L_{\rm bol}>L_{\rm thr}) = \mathrm{HMF}(M_{\rm h})\int_{L_{\rm thr}}^{\infty}\mathrm{CLF}(L_{\rm bol}|M_{\rm h})\,\mathrm{d}L_{\rm bol}, \tag{4.3}$$

where HMF stands for the (sub)halo mass function. A comparison between the QHMF and the HMF can then return the value of the UV-luminous quasar duty cycle, $\varepsilon_{\rm QSO}$:

$$\varepsilon_{\rm QSO} = \frac{\int_{M_{\rm med}}^{\infty}\mathrm{QHMF}(M|L_{\rm bol}>L_{\rm thr})\,\mathrm{d}M}{\int_{M_{\rm med}}^{\infty}\mathrm{HMF}(M)\,\mathrm{d}M}. \tag{4.4}$$

The lower integration limit is set to the median value[5] of the QHMF, $M_{\rm med}$ (see, e.g., Ren et al. 2020). For more details on the parametrization employed for the CLF and the definition of the various quantities at play, we refer the reader to Sec. 2 in Pizzati et al. (2024a) and Sec. 2 and Appendix A in Pizzati et al. (2024b).

The framework developed in these works builds on large-volume, dark-matter-only cosmological simulations. In particular, Pizzati et al. (2024b) uses the new FLAMINGO-10k simulation (part of the FLAMINGO project, Schaye et al. 2023; Kugel et al. 2023), which evolves 10080^3 cold dark matter (CDM) particles and 5600^3 neutrino particles in a box size of $L = 2.8\,\mathrm{cGpc}$

[5]The median of the QHMF is defined as the halo mass $M_{\rm med}$ satisfying the relation $\int_{M_{\rm med}}^{\infty}\mathrm{QHMF}(M_{\rm h}) = 0.5\int_0^{\infty}\mathrm{QHMF}(M_{\rm h})$.

Table 4.1: Constraints (median values and 16th-84th percentiles) on the parameters describing the conditional luminosity function (CLF; eq. 4.1) of quasars at $z \approx 4$ and $z \approx 6$. Taken from Pizzati et al. (2024a,b).

Redshift	σ	$\log_{10} L_{\rm ref}$ [erg s^{-1}]	γ	$f_{\rm on}$ [%]
$z \approx 4$	$0.20^{+0.13}_{-0.08}$	$45.2^{+0.3}_{-0.3}$	$2.00^{+0.22}_{-0.23}$	51^{+32}_{-31}
$z \approx 6$	$0.55^{+0.37}_{-0.31}$	$46.45^{+0.79}_{-1.35}$	$3.17^{+0.32}_{-0.34}$	$3.9^{+21}_{-3.2}$

assuming the "3x2pt + all" cosmology from Abbott et al. (2022)[6]. The model includes subhalos, which are found using the upgraded Hierachical Bound-Tracing (HBT+) code (Han et al. 2012, 2018). Subhalo masses, $M_{\rm h}$, are defined as peak bound masses[7].

In the analysis performed in Pizzati et al. (2024a), we applied this framework to the quasar auto-correlation functions measured by Eftekharzadeh et al. (2015) ($z \approx 2.5$) and Shen et al. (2007) ($z \approx 4$) using wide-field spectroscopic surveys such as SDSS (York et al. 2000) and BOSS (Ross et al. 2013). In particular, we showed that the $z \approx 4$ clustering measurements of Shen et al. (2007) imply a characteristic host halo mass for quasars of $\log_{10} M_{\rm h}/{\rm M}_\odot \approx 13.3$, corresponding to a very large UV-luminous quasar duty cycle of $\varepsilon_{\rm QSO} = 33^{+34}_{-23}$ %. In Pizzati et al. (2024b), we extended the framework to interpret the quasar-galaxy cross-correlation function recently measured by Eilers et al. (2024) at $z = 6.25$. This work exploited the JWST NIRCam wide-field slitless spectroscopic mode to pick up [O III] emitting galaxies in quasars fields, and inferred the clustering of quasars by measuring the cross-correlation function between quasars and [O III] emitting galaxies in conjunction with the auto-correlation function of these galaxies. By simultaneously fitting these two quantities, Pizzati et al. (2024b) found a characteristic host mass for $z \approx 6$ quasars of $\log_{10} M_{\rm h}/{\rm M}_\odot \approx 12.5$, lower than the one found at $z \approx 4$ and in line with results at $z \approx 2.5$.

However, when converting these host halo masses into peak heights[8], $\nu(M_{\rm h})$ – which measure how rare the large-scale over-density fluctuations are in the original linear field – we find that quasar clustering measurements at $z \approx 4$ and $z \approx 6$ point to similar values of $\nu \approx 4 - 6$. This implies that high-z, UV-luminous quasars seem to live in similarly biased and over-dense

[6]The cosmology parameters are: $\Omega_{\rm m} = 0.306$, $\Omega_{\rm b} = 0.0486$, $\sigma_8 = 0.807$, $\rm H_0 = 68.1\ km\,s^{-1}\,Mpc^{-1}$, $n_{\rm s} = 0.967$; the summed neutrino mass is 0.06 eV.

[7]In practice, we compute the mass of each (sub)halo by summing up the mass of all its bound particles and consider the largest mass that a (sub)halo has had across cosmic history.

[8]The peak height $\nu(M_{\rm h}, z)$ is formally defined as $\nu(M_{\rm h}, z) = \delta_{\rm c}/\sigma(M_{\rm h}, z)$ – with $\delta_{\rm c} \approx 1.69$ being the critical linear density for spherical collapse and $\sigma^2(M_{\rm h}, z)$ the variance of the linear density field smoothed on a scale $R(M_{\rm h})$; we compute $\nu(M_{\rm h}, z)$ using the python package COLOSSUS (Diemer 2018, see Sec. 5 in Pizzati et al. 2024b).

environments, corresponding to $(4-6)\sigma$ peaks in the initial linear density field (see also, e.g., Costa 2024). Due to the rapid decline of the unobscured quasar number density with redshift (solid gray line in the right panel of Fig. 4.1), these similar environments lead to very different values for the quasar UV-luminous duty cycles at $z \approx 4$ and $z \approx 6$: while UV-luminous $z \approx 4$ quasars are sufficiently abundant to occupy a large fraction of the coeval $\nu \approx 4-6$ halos, at $z \approx 6$ quasars are so rare that the same occupation fraction drops by more than an order of magnitude, with an implied duty cycle of $\varepsilon_{\rm QSO} = 0.9^{+2.3}_{-0.7}\%$.

We report the inferred values of the parameters describing the CLF and the $L_{\rm c}(M)$ relation (eq. 4.1-4.2) at $z \approx 4$ and $z \approx 6$ in Table 4.1. Further discussion on the comparison between quasar clustering results at these two redshifts can be found in Sec. 5 of Pizzati et al. (2024b) (see also Eilers et al. 2024). We mention the caveat, however, that the strong clustering measured at $z \approx 4$ is rather surprising and it is yet to be fully accounted for by any evolutionary models of quasar activity (Pizzati et al. 2024a, and references therein). Additionally, several other studies (e.g., Timlin et al. 2018; He et al. 2018; García-Vergara et al. 2019) have also attempted to measure quasar clustering at $z \approx 4$, challenging the exceptionally strong clustering inferred by Shen et al. (2007). Nevertheless, the Shen et al. (2007) measurement remains the most robust, as it is based on a large sample of spectroscopically-selected quasars. Future spectroscopic surveys (such as DESI, Yang et al. 2023) will further refine these measurements and provide more stringent constraints on the quasar auto-correlation function up to $z \approx 5$. Here, we take the Shen et al. (2007) result at face value, but stress the fact that our conclusions for $z \approx 4$ and $z \approx 6$ are completely independent.

In Fig. 4.2, we show the QHMFs obtained by our model at $z \approx 4$ and $z \approx 6$, together with HMFs at the respective redshifts. As discussed above, the QHMF can be obtained only once a bolometric luminosity threshold for quasars has been set. Both quasar clustering measurements on which our work is based (Shen et al. 2007; Eilers et al. 2024) focus on very bright unobscured quasars with $L_{\rm bol} \approx 10^{47}\,{\rm erg\,s^{-1}}$, with the work of Shen et al. (2007) extending down to slightly fainter objects of $L_{\rm bol} > 10^{46.7}\,{\rm erg\,s^{-1}}$. For consistency (see also Appendix D of Pizzati et al. 2024b), we show our $z \approx 6$ QHMF results setting the same bolometric luminosity threshold employed by Shen et al. (2007) at $z \approx 4$ (i.e., $L_{\rm bol} = 10^{46.7}\,{\rm erg\,s^{-1}}$). The QHMFs obtained in this way are plotted in Fig. 4.2 with red ($z \approx 6$) and green ($z \approx 4$) lines, and labeled as "bright quasars" as they only refer to the bright end of the unobscured quasar population.

Fainter, unobscured quasars are found at both $z \approx 4$ and $z \approx 6$ down to $L_{\rm bol} \approx 10^{45.3}\,{\rm erg\,s^{-1}}$ (Akiyama et al. 2018; Kulkarni et al. 2019; Matsuoka et al. 2022). However, the clustering of this fainter population is still largely unconstrained in the high-z Universe. A first attempt at measuring the clustering of $z \approx 6$ faint quasars was made by Arita et al. (2023): despite the

Table 4.2: Constraints (median values and 16th-84th percentiles) on the UV-luminous active fraction $f_{\rm on,UV}$ (coming from clustering measurements; see Pizzati et al. 2024a,b) and on the obscured:unobscured ratio for LRDs, $r_{\rm obsc}$ (from abundance arguments; see Sec. 4.2), at $z = 4$ and $z = 6.25$. The product $f_{\rm on,UV} r_{\rm obsc}$ exceeds unity at both redshifts, which is unphysical. In the last columns, we also report the median mass, $M_{\rm med}^{\rm (faint)}$ for the halos hosting faint unobscured quasars (see Fig. 4.2) and the number density of halos above this mass, $n_{\rm h}(> M_{\rm med}^{\rm (faint)})$, to be compared with the LRD number density of $\Phi_{\rm LRD} \approx 1.3 \times 10^{-5}\,{\rm cMpc}^{-3}$.

Redshift	$\log_{10} f_{\rm on,UV}$	$\log_{10} r_{\rm obsc}$	$\log_{10} f_{\rm on,UV} \cdot r_{\rm obsc}$	$f_{\rm on,UV} \cdot r_{\rm obsc}$	$\log_{10} M_{\rm med}^{\rm (faint)}/{\rm M}_\odot$	$n_{\rm h}(> M_{\rm med}^{\rm (faint)})$	$\Phi_{\rm LRD}/n_{\rm h}(> M_{\rm med}^{\rm (faint)})$
$z = 6.25$	$-1.40^{+0.83}_{-0.74}$	2.9 ± 0.5	$1.5^{+1.0}_{-0.9}$	32^{+284}_{-28}	≈ 12.15	$\approx 7.2 \times 10^{-7}\,{\rm cMpc}^{-3}$	≈ 18
$z = 4$	$-0.29^{+0.21}_{-0.41}$	1.3 ± 0.3	$1.0^{+0.4}_{-0.5}$	10^{+15}_{-7}	≈ 12.76	$\approx 1.6 \times 10^{-6}\,{\rm cMpc}^{-3}$	≈ 8.1

large uncertainties at play, these authors find a relatively large characteristic host halo mass of $M_{\rm h} = 7^{+11}_{-6} \times 10^{12}\,{\rm M}_{\odot}$ (but see Appendix C of Pizzati et al. 2024b, where it is shown that different assumptions on the quasar correlation function make these constraints much weaker). The relatively large inferred host mass for the faint quasar population would be in line with results at lower redshift ($z \lesssim 2.5$), which generally predict little to no dependence of quasar clustering with bolometric luminosity (e.g., Shen et al. 2009; Eftekharzadeh et al. 2015).

As our model is based on an empirical relation between quasar luminosities and (sub)halo masses, it can be used to predict the clustering of faint unobscured quasars at high redshift. With light-colored lines in Fig. 4.2, we plot the predictions for the "faint quasars" QHMFs at the two redshifts of interest. These QHMFs are obtained by lowering the bolometric luminosity threshold, $L_{\rm thr}$ in eq. 4.3, down to $L_{\rm bol} = 10^{45.5}\,{\rm erg\,s^{-1}}$. We note that such a low bolometric luminosity threshold implies that the results are sensitive to the relation between *faint* quasar luminosities and host halo masses. This relation is based on the extrapolation of our CLF parametrization down to low $L_{\rm bol}$, and it currently lacks support by constraints on the clustering of faint unobscured quasars. However, our fitting framework matches the unobscured QLF over the entire range of magnitudes, from the very bright to the very faint end, with a minimal number of parameters. Therefore, while faint quasar clustering studies will ultimately test our predictions, the QHMFs shown in Fig. 4.2 for faint quasars represent our best knowledge of how faint quasars populate the host halo mass spectrum, and are informed by our current understanding of unobscured quasar demographics.

At $z \approx 6$ (left panel of Fig. 4.2), we predict that the "faint quasars" QHMF peaks at $\log_{10} M_{\rm h}/{\rm M}_{\odot} \approx 12.15$, with a rather large spread in the host mass distribution (0.5 dex at 1 standard deviation). This implies a very mild dependence of clustering on bolometric luminosity, as a change of ≈ 1 dex in $L_{\rm bol}$ results in a change of ≈ 0.3 dex in the median of the host mass distribution, $M_{\rm med}$. This mild dependence is driven by two factors: a steep $L_{\rm bol} - M_{\rm h}$ relation and a large scatter around this relation[9]. These results are in broad agreement with clustering studies at low redshift, which find little to no dependence of clustering strength on luminosity (Croom et al. 2005; Myers et al. 2006; Shen et al. 2009) and attribute that to a large scatter in quasar luminosities at fixed halo mass (e.g., Adelberger & Steidel 2005; Lidz et al. 2006).

The strong clustering measured for bright quasars at $z \approx 4$ implies a slightly different dependence of quasar clustering on luminosity, with ≈ 1 dex in $L_{\rm bol}$ corresponding to ≈ 0.5 dex in $M_{\rm med}$. Such a luminosity dependence is a consequence of the large duty cycle measured for bright quasars: if

[9]The slope of the $L_{\rm bol} - M_{\rm h}$ relation and its scatter are directly constrained by a combination of the quasar clustering strength and the *shape* of the quasar luminosity function (Pizzati et al. 2024a).

these quasars occupy a large fraction of the available massive halos, fainter quasars will inevitably need to reside in less massive hosts. In practice, this is achieved in our model with a small predicted scatter for the $z \approx 4$ $L_{\rm bol} - M_{\rm h}$ relation (also found by White et al. 2008; Wyithe & Loeb 2009; Shankar et al. 2010b). The slightly different dependence of clustering on luminosity at the two redshifts considered, while interesting, has little impact on the conclusions presented in this work: at both redshifts, faint quasars also live in massive halos corresponding to highly biased environments which trace back to rare $\approx 4\sigma$ fluctuations in the linear density field.

4.3.2 Connecting the UV-luminous duty cycle to the AGN obscured population

Having described current constraints on the duty cycle and host mass distribution of UV-luminous, unobscured quasars, we turn our attention to the large population of LRDs/obscured AGN discussed in Sec. 4.2. The most general question connected to this obscured high-z population is how it fits into our understanding of SMBH accretion/AGN activity across the history of the Universe. In this context, determining whether LRDs and UV-selected quasars are drawn from the same population of dark matter halos can offer key insights into the nature of these sources. According to AGN unification models (e.g., Antonucci 1993; Padovani et al. 2017), the diversity of AGN emission across the electromagnetic spectrum can be entirely explained by a viewing-angle effect: the intrinsic emission from a quasar/AGN varies for different lines of sight because of, e.g., dust and gas obscuration. The natural consequence of this model is that all types of AGN (irrespective of their observed SEDs) share the same *intrinsic* properties, such as the bolometric luminosity, SMBH mass, and host halo mass distributions. Hence, if LRDs fit into this AGN unification picture, we expect them to reside in the same halos as comparably luminous UV-selected quasars. However, several studies at low z have challenged this AGN unification scenario by showing that obscured (type-2 or reddened type-1) quasars live in different dark matter halos than those of UV-luminous, type-1 quasars (e.g., Hickox et al. 2011; Allevato et al. 2014; Petter et al. 2023; Córdova Rosado et al. 2024). According to these studies, obscured quasars/AGN represent a different stage in the co-evolution between accreting SMBHs and their host galaxies/halos. Analogously, LRDs could also represent a different evolutionary phase in the accretion history of SMBHs. If that is the case, the host halo mass distribution of LRDs could be different than the one of unobscured quasars, even when matching their bolometric luminosities and SMBH masses. An obvious consequence of this hypothesis is that LRDs would be described by very different scaling relations (e.g., SMBH mass-halo/galaxy mass) than those in place for UV-luminous quasars, as identical SMBH masses would correspond to very different host halo/galaxy masses.

In this work, we point out that an indirect answer to whether LRDs and UV-selected quasars reside in the same dark matter halos comes from current constraints on the clustering of quasars at $z \approx 4-6$ (Sec. 4.3.1). From these constraints, we conclude that LRDs and unobscured quasars cannot be drawn from the same host halo distribution. Hence, their different SED properties reveal fundamental differences in their scaling relations. Our argument is simple: clustering measurements determine the host mass distribution of unobscured quasars; if LRDs followed the same distribution, the large obscured fraction derived in Sec. 4.2 implies that LCDM cosmology would not produce enough halos at these masses to accommodate this abundant population.

The argument can be visualized in Fig. 4.2: using dark (light) purple lines, we show the QHMFs of bright (faint) quasars scaled up by the obscured:unobscured ratios, $r_{\rm obsc}$, determined in Sec. 4.2 (plotted with colored arrows for reference). These obscured ratios are independent of bolometric luminosities, and increase rapidly with redshift from $r_{\rm obsc} \approx 20^{+20}_{-10} : 1$ at $z = 4$ to $r_{\rm obsc} \approx 815^{+1600}_{-545} : 1$ at $z = 6.25$. By multiplying the QHMF by $r_{\rm obsc}$, we are effectively computing the host mass distribution for LRDs/obscured AGN under the hypothesis that they reside in the same kind of halos as UV-luminous quasars. At both $z = 6.25$ (left panel in Fig. 4.2) and $z = 4$ (right), the host halo mass distributions for LRDs exceed the respective halo mass functions (HMFs). This is unphysical: cosmology sets hard (and well-constrained) limits on the number of (sub)halos that are available as quasar hosts as a function of mass, and the LRD number densities appear to be incompatible with these limits[10].

We can quantify this by considering the UV-luminous active fraction, $f_{\rm on,UV}$, which is a parameter in our CLF model (see eq. 4.1) and is closely related to the UV-luminous duty cycle (Pizzati et al. 2024a). The parameter $f_{\rm on,UV}$ represents the fraction of SMBHs that are actively accreting and unobscured at the same time. If we multiply this UV-luminous active fraction by the obscured:unobscured ratio $r_{\rm obsc}$, we are effectively computing an "obscured" active fraction (i.e., the fraction of halos hosting actively accreting LRDs/obscured AGN). The physical limit set by the number of available sub(halo) hosts can be then rephrased as $f_{\rm on,UV} \cdot r_{\rm obsc} < 1$. In Table 4.2, we report the values of $f_{\rm on,UV}$ and $r_{\rm obsc}$ and of their product at the two redshifts of interest, $z = 6.25$ and $z = 4$. We find that, despite the large uncertainties at play, these products are significantly larger than unity, with a value of $f_{\rm on,UV} \cdot r_{\rm obsc} \approx 10$ at $z = 4$ and $f_{\rm on,UV} \cdot r_{\rm obsc} \approx 36$ at $z = 6.25$. Coming back to the visual representation in Fig. 4.2, the product

[10]Note that this argument is valid only for a maximum occupation fraction of unity. Given that we model the distribution of all subhalos, however, it is natural to assume that each subhalo can host at most one accreting SMBH at its center.

$f_{\rm on,UV} \cdot r_{\rm obsc}$ represents the maximum ratio between the scaled-up QHMFs (purple lines; see also eq. 4.3) and the HMFs (dashed lines).

An even simpler way to present this argument is to consider the median mass values, $M_{\rm med}$, for, e.g., the faint-quasar QMHFs (Fig. 4.2, light-colored lines). In the right-hand columns of Table 4.2, we report these $M_{\rm med}^{\rm (faint)}$ values at the two redshifts of interest, together with the number density of halos above these mass thresholds, $n_{\rm h}(> M_{\rm med}^{\rm (faint)})$. When compared to the number density of LRDs, $\Phi_{\rm LRD} \approx 1.3 \times 10^{-5}\,{\rm cMpc}^{-3}$ (which is approximately constant with redshift; see Sec. 4.2), these number densities are a factor of ≈ 8.1 (≈ 18) smaller at $z = 4$ ($z = 6.25$). This confirms the fact that LRDs are simply too numerous to live in the same (sub)halos as UV-luminous quasars. As discussed in Sec. 4.3.1, these halo masses correspond to similar environments at $z \approx 4$ and $z \approx 6$ (i.e., $(4-6)\sigma$ peaks in the linear density field; Fig. 4.2). Since the number density of these environments is roughly constant with redshift (e.g., Tinker et al. 2008) (and so is $\Phi_{\rm LRD}$), LRDs outnumber their candidate host halos by similar factors at the two redshifts considered.

As a final note, we point out that our results are valid for any values of the quasar bolometric luminosities. Yet, in Sec. 4.2 and 4.3.1 we considered bright ($L_{\rm bol} > 10^{46.7}\,{\rm erg\,s^{-1}}$) and faint ($10^{45.5}\,{\rm erg\,s^{-1}} < L_{\rm bol} < 10^{46.7}\,{\rm erg\,s^{-1}}$) quasars separately because their properties are constrained differently. In particular, the abundance of obscured AGN is better constrained at faint bolometric luminosities by the large sample of LRDs with $L_{\rm bol} \approx 10^{46}\,{\rm erg\,s^{-1}}$; the evidence for an analog obscured population at large bolometric luminosities is instead only tentative (Sec. 4.2). On the other hand, the clustering of bright unobscured quasars has been directly measured (Sec. 4.3), but the QHMF and duty cycle for the faint quasar population is solely based on the extrapolation of our model to fainter bolometric luminosities – which constrained to match the faint end of the QLF. For this reason, the results presented lead to different conclusions depending on the bolometric luminosities considered. If a large obscured population is indeed present at $L_{\rm bol} \approx 10^{47}\,{\rm erg\,s^{-1}}$, then this is already in direct conflict with with constraints on the host masses and duty cycle of bright unobscured quasars (Shen et al. 2007; Eilers et al. 2024). A measurement of quasar clustering at the faint end of the QLF ($L_{\rm bol} \gtrsim 10^{45.5}\,{\rm erg\,s^{-1}}$), on the other hand, would provide support for our predictions for the properties of faint unobscured quasars, and will make it possible to directly compare the properties of UV-luminous quasars and LRDs at the same bolometric luminosities.

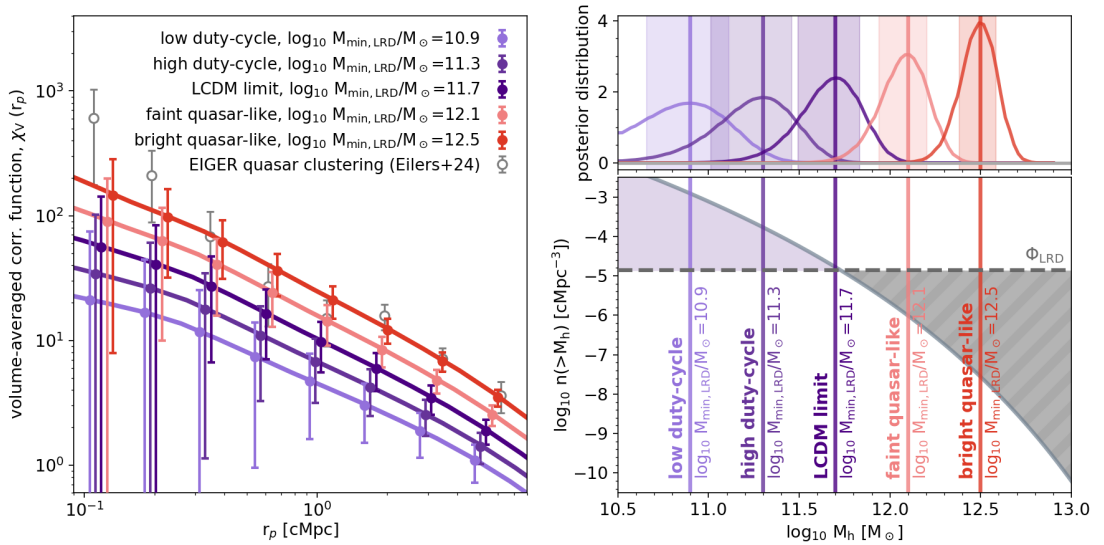

Figure 4.3: *Left:* Mock measurements (colored data points) for the LRD-galaxy cross-correlation functions obtained for different values of the minimum host mass for LRDs, $M_{\rm min,LRD}$. The measurements are obtained by putting together 10 different LRD fields, and extracting galaxy counts by setting a minimum host mass for galaxies (i.e., [O III] emitters) of $M_{\rm min,OIII} = 10^{10.56}\,{\rm M}_\odot$ and a background galaxy number density of $n_{\rm OIII} = 7.84 \times 10^{-4}\,{\rm cMpc}^{-3}$. The theoretical predictions for these cross-correlation functions are coming from the model of Pizzati et al. (2024b) and are shown as solid colored lines. Error bars are computed by assuming Poisson uncertainties on the galaxy number counts. Gray points refer to the UV-luminous quasar-galaxy cross-correlation function measurements from the EIGER survey (Eilers et al. 2024). *Top right:* Mock inference analysis for the LRD-galaxy cross-correlation function measurements, as a function of the minimum host LRD mass, $M_{\rm min,LRD}$. Values of $M_{\rm min,LRD}$ considered for the mock measurements are color-coded as in the other panels. The posterior distributions are obtained by computing the agreement between the mock measurements and the theoretical models for different minimum host LRD masses. Shaded regions show the 16th and 84th percentiles of their respective posterior distributions. *Bottom right:* Number density of $z = 6.25$ halos above $M_{\rm h}$, $n_{\rm h}(> M_{\rm h})$, as a function of halo mass $M_{\rm h}$ (solid gray line). The values of $M_{\rm min,LRD}$ considered in the analysis are highlighted with colored vertical lines. The dashed horizontal line corresponds to the LRD number density, $\Phi_{\rm LRD}$. The purple (gray) shaded area shows the region for which $M_{\rm h} < 10^{11.7}\,{\rm M}_\odot$ ($M_{\rm h} > 10^{11.7}\,{\rm M}_\odot$). In the purple region, $\Phi_{\rm LRD} < n_{\rm h}(> M_{\rm h})$ and hence the number of LRDs is less than the number of host halos available, whereas the gray region is unphysical as LRDs are too abundant for the number of host (sub)halos (assuming a maximum occupation fraction of unity).

4.4 The host mass and duty cycle of little red dots: a mock analysis

The indirect arguments presented in the previous section suggest that LRDs cannot live in the same dark matter halos as unobscured UV-luminous quasars, and hence – provided their bolometric luminosities are correctly estimated – they may constitute a fundamentally different population of accreting SMBHs. How do we determine this new population's host halo masses and duty cycle? In this section, we argue that this can be done using current (and upcoming) JWST observations.

Existing JWST programs such as EIGER (Kashino et al. 2023; Eilers et al. 2024) and ASPIRE (Wang et al. 2023) have already shown that the clustering of luminous, UV-selected quasars can be effectively measured using JWST NIRCam slitless spectroscopy to study the distribution of [O III] line emitting galaxies in the neighboring regions of the quasars. The same strategy can be applied to any other population of objects: the cross-correlation between this population and [O III] line emitters at a certain redshift can be measured, and the clustering of this said population can be inferred by simultaneously constraining the auto-correlation function of the [O III] line emitters.

In the following, we examine a simple proof-of-concept analysis that aims to measure the clustering of LRDs using JWST[11]. We focus here on $z = 6.25$, which is the redshift at which the clustering of UV-luminous quasars with [O III] emitters has already been measured by the EIGER survey (Eilers et al. 2024, see also Sec. 4.3.1). Following Eilers et al. (2024) (see also, e.g., Kaiser 1984; Martini & Weinberg 2001; Haiman & Hui 2001), we postulate that LRDs inhabit a fraction of all the (sub)halos whose mass is larger than some minimum mass threshold, $M_{\rm min,LRD}$[12]. This fraction is equal to the LRD duty cycle, $\varepsilon_{\rm LRD}$, and can be determined by comparing the LRD number density ($\Phi_{\rm LRD}$ in Sec. 4.2) to the abundance of halos with $M_{\rm h} > M_{\rm min}$. We note that we only consider LRDs with quasar-like bolometric luminosities (i.e., with the same bolometric luminosities as faint $z \approx 6$ quasars, $L_{\rm bol} > 10^{45.5}\,{\rm erg\,s^{-1}}$), as we are interested in matching LRDs and UV-luminous quasars in $L_{\rm bol}$ space.

[11] An alternative approach would be to directly measure the auto-correlation function of LRDs. Even though LRDs have a relatively high number density, however, measuring an autocorrelation function would require very large samples that are challenging to obtain given the small FoV of JWST.

[12] In other words, we do not model the LRD host mass distribution parametrically as described in Sec. 4.3 for unobscured quasars, but we assume that such a distribution can be obtained by rescaling the HMF above the minimum mass threshold $M_{\rm min,LRD}$. A more sophisticated parametrization would result in large degeneracies in the parameter space that could not be resolved by clustering measurements alone (e.g., Pizzati et al. 2024a; Muñoz et al. 2023).

We consider five different values of the minimum host (sub)halo mass for LRDs: $\log_{10} M_{\rm min,LRD}/{\rm M}_\odot = 10.9, 11.3, 11.7, 12.1, 12.5$. In the bottom right panel of Fig. 4.3, we put these values into context by showing the number density of $z = 6.25$ halos above $M_{\rm h}$, $n_{\rm h}(> M_{\rm h})$, as a function of halo mass (solid grey line); we highlight the values of $M_{\rm min,LRD}$ considered with colored vertical lines. By comparing the LRD number density ($\Phi_{\rm LRD}$, dashed horizontal line) to the integrated halo mass function ($n_{\rm h}(> M_{\rm h})$) for different minimum halo masses, we can directly relate the abundance of LRDs to that of available host dark matter halos. We find that the number of LRDs equals the number of host halos (i.e., the duty cycle is equal to unity) for a minimum host mass of $M_{\rm min,LRD} \approx 10^{11.7}\,{\rm M}_\odot$. Assuming that there can be only one LRD per (sub)halo, values of $M_{\rm min,LRD}$ above this threshold mass are unphysical. Values significantly lower than this threshold, on the other hand, imply a low duty cycle for LRDs, as only very few (sub)halos host LRDs at any given time.

Based on this discussion, we refer to the five different $M_{\rm min,LRD}$ cases considered in the following way (see Fig. 4.3): "low duty-cycle" ($M_{\rm min,LRD}/{\rm M}_\odot = 10^{10.9}\,{\rm M}_\odot$), "high duty-cycle" ($M_{\rm min,LRD}/{\rm M}_\odot = 10^{11.3}\,{\rm M}_\odot$), "LCDM limit" ($M_{\rm min,LRD}/{\rm M}_\odot = 10^{11.3}\,{\rm M}_\odot$), "faint quasar-like" ($M_{\rm min,LRD}/{\rm M}_\odot = 10^{12}\,{\rm M}_\odot$), "bright quasar-like" ($M_{\rm min,LRD}/{\rm M}_\odot = 10^{12.5}\,{\rm M}_\odot$). The first case ("low duty-cycle") corresponds to a duty cycle of $\varepsilon_{\rm LRD} \approx 1\%$, which is close to the duty cycle measured by Pizzati et al. (2024b) for UV-luminous quasar activity at the same redshift. In the second case, the implied LRD duty cycle increases to $\varepsilon_{\rm LRD} \approx 10\%$. The third case corresponds to the physical limit of a duty cycle of $\approx 100\%$. The last two cases, instead, would imply a duty cycle above unity, and correspond to host masses characteristic of UV-luminous quasars. Based on the discussion of Sec. 4.3.1, we associate the case $M_{\rm min,LRD} = 10^{12.1}\,{\rm M}_\odot$ to faint ($L_{\rm bol} \gtrsim 10^{45.5}\,{\rm erg\,s^{-1}}$) quasars – which have the same $L_{\rm bol}$ as LRDs –, while the larger mass of $M_{\rm min,LRD} = 10^{12.5}\,{\rm M}_\odot$ is close to the one found for luminous ($L_{\rm bol} \approx 10^{47}\,{\rm erg\,s^{-1}}$) unobscured quasars by Eilers et al. (2024).

The question we want to address here is whether we can use clustering measurements based on JWST slitless spectroscopy data to distinguish between these different $M_{\rm min,LRD}$ cases. We consider the following mock setup: JWST/NIRCam grism is used to image 10 different LRD fields. The distribution of [O III] line emitters in these fields can be employed to measure a LRD-galaxy cross-correlation function, from which the host mass and duty cycle of LRDs can be determined by exploiting the constraints on the galaxy-galaxy auto-correlation function (Eilers et al. 2024, Huang et al., in prep.).

In practice, we use the framework developed in Pizzati et al. (2024b), which outputs the cross-correlation function of any populations of objects that

are tracers of the underlying distribution of dark matter halos[13]. We employ this model to predict the LRD-galaxy cross-correlations for the different values of $M_{\rm min,LRD}$. [O III] line emitters are assumed to live in halos with a fixed threshold mass of $M_{\rm min,OIII} = 10^{10.56}\,{\rm M}_\odot$, which is set according to the results of Eilers et al. (2024) (see also Huang et al., in prep.). Based on these cross-correlation functions, we generate mock measurements by computing the expected number of galaxies as a function of the projected distance in each LRD field. The expected galaxy counts are obtained by setting a background galaxy number density of $n_{\rm OIII} = 7.84 \times 10^{-4}\,{\rm cMpc}^{-3}$, which is obtained by integrating the [O III] emitter luminosity function of Matthee et al. (2023) down to the threshold luminosity of $L_{\rm OIII,5008} = 10^{42}\,{\rm erg\,s^{-1}}$. We put together the 10 mock LRD fields and we compute the volume-averaged cross-correlation function, χ_V, by projecting the galaxy 3-d distributions over a comoving distance of $\pi_{\rm max} = 9.8\,{\rm cMpc}$, corresponding to a line-of-sight velocity of 1000 $\rm km\,s^{-1}$ at the redshift considered.

In the left panel of Fig. 4.3, we show the mock LRD-galaxy cross-correlation functions for different values of $M_{\rm min,LRD}$. We also show for reference the UV-luminous quasar-galaxy cross-correlation function measured by Eilers et al. (2024) by putting together 4 different quasars fields from the EIGER survey (Kashino et al. 2023). We note that, as also done in Eilers et al. (2024), the error bars we show are computed by considering only the contribution of Poisson uncertainties on the number counts. Other contributions to the error budget, such as cosmic variance or possible correlations between different data points, are neglected in this work and will be analyzed in a forthcoming study (Huang et al., in prep.).

The precision of our inference analysis is shown in the bottom left panel of Fig. 4.3. These posterior distributions are obtained by fitting the mock data with the LRD-galaxy cross-correlation function models obtained by varying the LRD mass threshold parameter, $\log_{10} M_{\rm min,LRD}/{\rm M}_\odot$. For each of these models, we compute the value of the χ^2 and plot in Fig. 4.3 the quantity $\exp(-\chi^2/2)$ (normalized to unity). By looking at the different posterior distributions, we learn that by putting together 10 LRD fields we can already constrain the values of $M_{\rm min,LRD}$ (and hence the characteristic host mass of LRDs) with an uncertainty of $\approx 0.1 - 0.3$ in $\log_{10} M_{\rm h}$. The posteriors are narrower and more peaked for larger $M_{\rm min,LRD}$. This follows from the fact that high-mass halos are more strongly clustered, and hence the clustering signal is stronger for large $M_{\rm min,LRD}$ (left panel of Fig. 4.3). In all cases considered, the uncertainty in $M_{\rm min,LRD}$ is small enough that,

[13]We use the FLAMINGO-10k large-volume cosmological simulation (Sec. 4.3) to build an analytical model for the cross-correlation function of any sets of halos with masses M_j and M_k, $\xi_h(M_j, M_k; r)$. An appropriate weighting scheme can then return the cross-correlation function between two different halo tracer populations. For more details on the model and the cosmological simulation employed, we refer the reader to Pizzati et al. (2024b).

in principle, it could be possible to tell apart the different scenarios. A larger number of LRD fields would be necessary, however, to reduce the uncertainties on $M_{\rm min,LRD}$ even further, and pinpoint its value even for the case of low $M_{\rm min,LRD}$.

The discussion presented here shows that, by measuring how galaxies cluster in LRD fields, it is indeed possible to determine whether LRDs live in the same dark matter halos as unobscured quasars (in agreement with, e.g., the AGN unification framework) or whether they are hosted by more common and less-biased environments, as it appears to be necessary given their large number density (see Sec. 4.3). In this latter case, measuring the host mass distribution of LRDs would also provide a way to quantify their duty cycle ($\varepsilon_{\rm LRD}$), which is a fundamental quantity that can help us to shed light on the accretion history of these enigmatic objects. A large value of $\varepsilon_{\rm LRD} \approx 10\%$ would suggest that LRDs have been actively accreting for a large fraction of cosmic time ($\gtrsim 100\,{\rm Myr}$), and hence – assuming a standard value for the radiative efficiency – that they would be able to build the relatively large black hole masses that have been inferred from their broad optical lines (up to $\gtrsim 10^8\,{\rm M}_\odot$; e.g., Greene et al. 2024; Kocevski et al. 2024). In particular, an accretion timescale of $\gtrsim 100\,{\rm Myr}$ corresponds to $\gtrsim 2t_{\rm S}$, where $t_{\rm S} \approx 45\,{\rm Myr}$ is the Salpeter time for exponential black hole mass growth (Salpeter 1964). This implies that LRDs are detectable above the observational luminosity threshold for at least a few Salpeter times, which is expected if the survey spans about one order of magnitude in luminosity. We point out that, for the same reason, large duty cycles of $\gtrsim 50\%$ are not to be expected, because they would imply that almost all LRDs shine above the observational threshold for a time that is much longer than the Salpeter timescale. Considering again a survey spanning about one order of magnitude in luminosity, a standard Eddington-limited growth that remains above the observational threshold for a time $t \gg 2t_{\rm S}$ would result in black holes that grow much more than one order of magnitude, and hence end up being more massive than what is actually observed. For this reason, while the threshold mass of $M_{\rm min,LRD} \approx 10^{11.7}\,{\rm M}_\odot$ represents a limit set by cosmological constraints on the number of available (sub)halos, black hole formation physics suggests an even more stringent limit on $M_{\rm min,LRD}$: if we require $\varepsilon_{\rm LRD} \lesssim 30\%$, this implies that $M_{\rm min,LRD}$ needs to be lower than $\approx 10^{11.5}\,{\rm M}_\odot$.

A very low LRD duty cycle $\varepsilon_{\rm LRD} \lesssim 1\%$, on the other hand, would also be puzzling, as it would raise the question of how to reconcile the large black hole masses measured for LRDs with their inherently sporadic activity. This is the same problem that has been brought up for the high-z UV-luminous quasar population, for which different methods generally infer low values of the quasar duty cycle and/or quasar lifetime (e.g., Khrykin et al. 2016, 2019; Eilers et al. 2018, 2020; Davies et al. 2018, 2019, 2020; Worseck et al. 2016, 2021; Ďurovčíková et al. 2024; Eilers et al. 2024) that appear to be in

tension with their large, $\gtrsim 10^9\,\mathrm{M}_\odot$ black hole masses. A possible solution to explain a low value of the duty cycle is super-Eddington accretion: if accretion on black holes takes place in short, radiatively inefficient bursts, then we expect a lower $\varepsilon_{\mathrm{LRD}}$ because the Salpeter timescale for black hole accretion becomes shorter than $\approx 45\,\mathrm{Myr}$. Interestingly, several studies have invoked super-Eddington accretion to explain the puzzling SED features of LRDs (e.g., Greene et al. 2024; Pacucci & Narayan 2024; Lambrides et al. 2024a). Measuring the clustering of LRDs and inferring their duty cycle would provide an independent way to support these claims.

Finally, if bright LRDs have large black hole masses ($\gtrsim 10^8\,\mathrm{M}_\odot$) but live in much smaller halos than UV-selected quasars, they need to obey fundamentally different scaling relations. Constraining the clustering of LRDs would provide insights into these relations: the lower the mass of the host halos, for instance, the more overmassive LRDs need to be with respect to the black hole mass-halo mass relation holding for unobscured quasars. We can also cast this in terms of the black hole mass-stellar mass relation – which has been extensively discussed in the recent literature (e.g., Pacucci et al. 2023; Yue et al. 2024a) – by converting halo masses to stellar masses using the relation provided by Behroozi et al. (2019). We find that halo masses in the range $M_{\mathrm{h}} \approx 10^{11} - 10^{11.5}\,\mathrm{M}_\odot$ correspond – at the redshift of interest – to stellar masses of $M_\star \approx 10^{8.4} - 10^{9.4}\,\mathrm{M}_\odot$. This implies that, assuming black hole mass measurements are not significantly overestimated, LRDs are highly overmassive with respect to the coeval black hole mass-stellar mass relation, as the ratio between black hole and galaxy masses would be in the range $\approx 10\% - 100\%$ (see also, e.g., Durodola et al. 2024).

4.5 Discussion and summary

In this work, we have examined how the new population of Little Red Dots (LRDs) revealed by JWST compares to the one of UV-luminous quasars that have been studied for decades using wide-field spectroscopic surveys (e.g. Fan et al. 2023). The basic observational evidence on which our work is based, is that a large fraction of LRDs exhibits broad emission lines in their spectra, whose properties directly point to the presence of AGN that are (at least partially) responsible for the observed emission (Greene et al. 2024; Kocevski et al. 2024). This, together with their very red colors at optical wavelengths, has led to the interpretation that LRDs could be standard, UV-luminous type-1 quasars whose radiation is (partially) obscured by intervening dust and gas. By correcting for the effects of this obscuration, it is possible to use broad lines to estimate the bolometric luminosities of the SMBHs accreting at the center of LRDs. Several works (e.g., Greene et al. 2024; Kokorev et al. 2024a; Akins et al. 2024) have shown that such (unattenuated) bolometric

luminosities extend up to $\approx 10^{46} - 10^{47}\,\mathrm{erg\,s^{-1}}$, well in the range that is characteristic of unobscured, type-1 quasars (Fig. 4.1, left panel).

Yet, the abundances of LRDs and UV-luminous quasars are remarkably different. In Fig. 4.1, we have directly compared the redshift evolution for the number density of UV-luminous quasars to the one for LRDs at the same bolometric luminosities. It is well-established that the abundance of unobscured quasars drops exponentially with increasing redshift (e.g., Richards et al. 2006; Schindler et al. 2023). Spectroscopic (Greene et al. 2024) and photometric (Kokorev et al. 2024a) surveys of LRDs, instead, find little to no evolution in their number density over a wide redshift range ($z \approx 4-8$), with an approximately constant value of $\Phi_{\mathrm{LRD}} \approx 1.3 \times 10^{-5}\,\mathrm{cMpc^{-3}}$ ($L_{\mathrm{bol}} > 10^{45.5}\,\mathrm{erg\,s^{-1}}$). By comparing the number density of LRDs to that of UV-luminous quasars as a function of redshift, we can estimate the obscured fraction of AGN implied by this LRD population. We infer a large and rapidly evolving obscured fraction, ranging from $\approx 20:1$ at $z \approx 4$ to $\approx 2300:1$ at $z \approx 7$.

While this obscured fraction is mostly constrained at the bolometric luminosities for which a significant overlap between LRDs and unobscured quasars is present (i.e., $L_{\mathrm{bol}} \approx 10^{45} - 10^{46}\,\mathrm{erg\,s^{-1}}$), we find tentative evidence for it to extend to even larger bolometric luminosities ($L_{\mathrm{bol}} \gtrsim 10^{47}\,\mathrm{erg\,s^{-1}}$). There are two arguments in support of this evidence: (a) photometric observations (Kokorev et al. 2024a) constrain the *shape* of the LRD bolometric luminosity functions to closely resemble that of UV-luminous quasars (Nida et al. 2020; Schindler et al. 2023; Matsuoka et al. 2023), implying an obscured fraction that is constant with bolometric luminosity (Fig. 4.1, left panel); (b) recent observations of the COSMOS field have revealed candidate radio-loud AGN at $z \approx 7-8$ that are obscured in the UV (Endsley et al. 2022, 2023; Lambrides et al. 2024b); the very large bolometric luminosities of these sources ($L_{\mathrm{bol}} \approx 10^{47}\,\mathrm{erg\,s^{-1}}$) together with the small FoV of the observations, implies an AGN obscured fraction that is consistent to the one inferred for bright LRDs.

The large abundance of LRDs/obscured AGN has implications for their host halo masses. If obscuration were solely a viewing angle effect (e.g., Antonucci 1993), then we would expect LRDs to reside in the same environments as (equally bolometrically bright) UV-luminous quasars. Two decades of quasar clustering studies have constrained the masses of the dark matter halos hosting UV-luminous quasars at $0 \lesssim z \lesssim 6$ to be in the range $M_h \approx 10^{12} - 10^{13.5}\,\mathrm{M_\odot}$ (e.g., Porciani et al. 2004; Croom et al. 2005; Porciani & Norberg 2006; Shen et al. 2007, 2009; Ross et al. 2013; Eftekharzadeh et al. 2015; Arita et al. 2023; Eilers et al. 2024), with little to no dependence on quasar luminosity (e.g., Adelberger & Steidel 2005; Porciani & Norberg 2006; Shen et al. 2009). Several models have been put forward to interpret this host mass range in physical terms (e.g., Hopkins et al. 2007b; Fanidakis et al. 2013; Caplar et al. 2015). Whatever the reason for these characteristic

host masses, it is striking that the number density of available host halos at these masses drops very quickly below the measured abundance of LRDs as redshift increases. At $z \approx 6$, for example, LRDs (with $L_{\rm bol} > 10^{45.5}\,{\rm erg\,s^{-1}}$) are $\approx 5\times$ more abundant than $10^{12}\,{\rm M_\odot}$ halos (Fig. 4.3, top right panel) and can occupy all halos above the threshold mass of $\tilde{M}_{\rm h} > 10^{11.7}\,{\rm M_\odot}$. This implies that at these redshifts the host masses of LRDs are likely lower than the ones of UV-luminous quasars, even when matching them in $L_{\rm bol}$ space.

In Fig. 4.2, we have presented a quantitative analysis of this argument at the two redshifts for which we have constraints on the clustering of bright ($L_{\rm bol} \approx 10^{47}\,{\rm erg\,s^{-1}}$), high-$z$ unobscured quasars: $z = 4$ (Shen et al. 2007) and $z = 6.25$ (Eilers et al. 2024). We used the model developed in Pizzati et al. (2024a,b) to measure the UV-luminous quasar host mass functions (QHMFs) at these two redshifts. While these QHMFs are well-constrained by clustering measurements only for the bright quasar population, we can extend them to also include the contribution of faint ($L_{\rm bol} \gtrsim 10^{45.5}\,{\rm erg\,s^{-1}}$) quasars by using the empirical quasar luminosity-halo mass relations obtained by Pizzati et al. (2024a,b). These relations are fit to the faint end of the quasar luminosity function, and hence they correctly reproduce the demographic properties of the faint quasar population. While we find minor differences in the luminosity dependence of the QHMFs at the two redshifts considered, we reach a general fundamental conclusion that is valid for faint and bright sources alike: *the dark matter halos hosting UV-luminous quasars at* $z \gtrsim 4$ *are too rare to accommodate the large number density of LRDs.*

What are the implications of these findings? If LRDs live in more common and hence less biased halos than those of unobscured quasars, then they may represent an intrinsically different population of accreting SMBHs arising in the early Universe. This population may be tracing a distinct phase in the co-evolutionary sequence of SMBHs and galaxies, similarly to what has been argued for type-2/reddened quasars at low redshifts (e.g., Allevato et al. 2014; Córdova Rosado et al. 2024). In this scenario, the scaling relations between, e.g., black hole and halo/galaxy host masses need to be intrinsically different for LRDs and standard unobscured quasars, because similar black hole masses correspond to very different halo (and hence galaxy) masses. In particular, LRDs likely host SMBHs that are overmassive with respect to the coeval black hole-halo/galaxy mass scaling relations for unobscured quasars. Another possibility that has been put forward by several independent works to explain the enigmatic features of LRD SEDs (e.g., Greene et al. 2024; Pacucci & Narayan 2024; Lambrides et al. 2024a) is that LRDs are accreting at rates that are larger than the critical Eddington limit. In this latter case, LRDs could represent the early stages of black hole accretion and growth that are predicted by many theoretical models of SMBH evolution (e.g., Trinca et al. 2023; Li et al. 2024; Lupi et al. 2024). Interestingly, this would have direct implications for the clustering of LRDs, because a low duty cycle

(that is necessary for super-Eddington accretion) would only be possible if LRDs lived in very low mass halos ($M_{\rm h} \approx 10^{11}\,{\rm M}_{\odot}$ at $z \approx 6$; Sec. 4.4).

Alternatively, these results may be telling us that key properties of LRDs, such as their bolometric luminosities and the relative contribution of the central AGN and the host galaxy to their observed SEDs, have yet to be properly characterized. Indeed, the assumption on which our discussion is based, is that LRDs have the same bolometric luminosities as high-z UV-luminous quasars ($L_{\rm bol} \approx 10^{45} - 10^{47}\,{\rm erg\,s^{-1}}$). Currently, the bolometric luminosities of LRDs are estimated by their (dereddened) broad emission lines or by fitting AGN templates to their SEDs. In both cases, the resulting $L_{\rm bol}$ hinge on the assumption that the rest-frame optical continuum is dominated by AGN light (see e.g., Akins et al. 2024). If the contribution of the host galaxy to the rest-frame optical continuum emission (and possibly broad lines; see, e.g., Baggen et al. 2024) is non-negligible, then the inferred black hole masses and bolometric luminosities could change significantly. Several puzzling features of LRDs, such as their X-ray weakness (Ananna et al. 2024; Yue et al. 2024b; Maiolino et al. 2024) and (possibly) the lack of a hot dust torus (Wang et al. 2024; Pérez-González et al. 2024; Akins et al. 2024; Iani et al. 2024) and UV variability (Kokubo & Harikane 2024), point to the fact that LRD bolometric luminosities could be vastly overestimated. The presence of an evolved stellar population dominating (part of) the rest-frame optical is also suggested by the detection of a Balmer break in some LRD spectra (e.g., Wang et al. 2024; Kokorev et al. 2024b, but see Inayoshi & Maiolino 2025), although the large densities and stellar masses required to match the observed LRD luminosities remain a significant challenge to a purely stellar interpretation of LRD SEDs (e.g., Greene et al. 2024; Akins et al. 2024, but see Baggen et al. 2024). Regardless of the exact AGN contribution to these SEDs, if LRDs are not as bright as standard, UV-luminous quasars then they would naturally reside in lower mass halos, and they could easily be accommodated in the large number of $z \gtrsim 6$ host halos with masses of $M_{\rm h} \approx 10^{11} - 10^{11.5}\,{\rm M}_{\odot}$.

In this work, we have primarily focused on the population of LRDs whose inferred SMBH masses and bolometric luminosities largely overlap with those of UV-luminous quasars. However, JWST has uncovered a much larger population of AGN with broad optical (Hα or Hβ) lines, which are not necessarily reddened at optical wavelengths and hence do not respect the LRD selection criteria. Interestingly, the abundance of these broad-line AGN are even larger than the ones of LRDs: Maiolino et al. (2024), Harikane et al. (2023), and Taylor et al. (2024) find the number densities for these sources to be in the range $10^{-3} - 10^{-5}\,{\rm cMpc^{-3}mag^{-1}}$ ($4 \lesssim z \lesssim 7$). The intrinsic bolometric luminosities and SMBH masses of these broad-line AGN (that are not reddened in the rest-frame optical) are not as extreme as the ones of LRDs/reddened AGN (e.g., Harikane et al. 2023; Taylor et al. 2024). However, these sources can still reach UV magnitudes of $M_{\rm UV} \approx -22$

and bolometric luminosities of $L_{\rm bol} \approx 10^{45.5}\,{\rm erg\,s^{-1}}$, which are close to the ones of the faintest UV-selected quasars known at $z \gtrsim 4$ (Matsuoka et al. 2022). Given their number densities, these broad-line AGN overshoot the extrapolation of the UV-selected quasar luminosity functions by factors that are comparable to (or even higher than) those found for LRDs (Sec. 4.2). Hence, similar arguments to the ones presented in our analysis apply to this larger AGN population: their abundance suggests that they live in halos that are likely less massive than those of comparably luminous UV-selected quasars, implying that they obey fundamentally different scaling relations. While a proper comparison between UV-selected quasars and JWST AGN is only possible for the LRD population with large inferred bolometric luminosities and SMBH masses, it is interesting to investigate the host mass distributions, duty cycles, and scaling relations of this larger population of faint broad-line AGN.

Ultimately, a measurement of the clustering of LRDs and other broad-line AGN will constrain such properties and test the conclusions that we have drawn in this work. Recent arguments on the clustering of these objects rely on single detections of AGN in close proximity (Lin et al. 2024; Tanaka et al. 2024), on spectroscopic detections of galaxies in a single LRD field (Schindler et al. 2024), and on cross-correlating photometrically-selected galaxies and LRDs (Arita et al. 2025). In this work (Fig. 4.3), we have shown that a convincing measurement of LRD clustering can be made by using JWST NIRCam/WFSS observations of several LRD fields to extract a cross-correlation function between LRDs and [O III] line-emitting galaxies (see also Matthee et al. 2024a for recent results based on a similar approach). We have suggested that, by putting together ≈ 10 different fields, it is possible to infer the characteristic host halo mass of LRDs with an accuracy of $\log_{10} M_{\rm h} \approx 0.1 - 0.3$. In order to perform this kind of measurement, one would need to observe several fields containing LRDs using a NIRCam grism filter covering the [O III] doublet. Interestingly, such observations already exist for a fraction of the broad-line AGN in the sample of Matthee et al. (2024b): JWST surveys such as CONGRESS (GO3577) and GTO4540/GTO4549 are performing NIRCam/WFSS observations of the GOODS-N and GOODS-S fields, which contains ≈ 10 broad-line AGN from the Matthee et al. (2024b) sample. So a first step towards determining the clustering of these enigmatic sources at $z \gtrsim 5$ is already feasible with current data. Future JWST programs will be able to deliver the same kind of observations for samples of LRDs with quasar-like inferred bolometric luminosities and SMBH masses. By comparing the host halo masses resulting from these measurements to the different scenarios discussed in Sec. 4.4, it will be possible to get fundamental insights into the properties of these objects.

At the same time, the clustering of the faint, UV-luminous quasar population at high redshifts is also largely unconstrained. By using the same

strategy and targeting faint quasar fields with NIRCam/WFSS, it will also be possible to determine their clustering and host masses. This would test our model predictions (Fig. 4.2) and determine the luminosity dependence of quasar clustering at high-z, effectively constraining the scaling relation between the quasar bolometric luminosity and the host halo mass. Even more importantly, it would create a benchmark to which the LRD population can be effectively compared, allowing us to investigate the nature of quasar activity and SMBH populations in the early Universe.

Acknowledgements

We acknowledge helpful conversations with Junya Arita, Jorryt Matthee, Jacob Shen, Fengwu Sun, Marta Volonteri, Minghao Yue, and Ben Wang. We are grateful to the FLAMINGO-10k team for making their simulation available. We are also grateful to the ENIGMA group at UC Santa Barbara and Leiden University for discussion of an early version of this manuscript. JFH and EP acknowledge support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 885301). JTS is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project number 518006966. FW acknowledges support from NSF Grant AST-2308258. This work is partly supported by funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860744 (BiD4BESt). This work used the DiRAC Memory Intensive service (Cosma8) at the University of Durham, which is part of the STFC DiRAC HPC Facility (www.dirac.ac.uk). Access to DiRAC resources was granted through a Director's Discretionary Time allocation in 2023/24, under the auspices of the UKRI-funded DiRAC Federation Project. The equipment was funded by BEIS capital funding via STFC capital grants ST/K00042X/1, ST/P002293/1, ST/R002371/1 and ST/S002502/1, Durham University and STFC operations grant ST/R000832/1. DiRAC is part of the National e-Infrastructure.

5 Tracing individual black hole growth histories and quasar lightcurves in an N-body Universe

Abstract

We present a new model for the evolution of supermassive black holes (SMBHs) and quasars across cosmic time. The framework builds on merger trees from the FLAMINGO large-volume dark-matter-only simulation, linking SMBH growth histories to those of their host halos through parametric prescriptions that capture both average trends and stochastic variability. SMBH accretion is modeled self-consistently and directly drives quasar activity, with the goal of reproducing the observed evolution of the luminous quasar population. Our model is designed to match three key observables: the bolometric quasar luminosity function (QLF), the conditional Eddington ratio distribution function (cERDF) at fixed bolometric luminosity, and the clustering of UV-luminous quasars. Despite residual discrepancies at the faint end of the QLF, in the mean of the cERDF, and in quasar clustering at $z \approx 4$, our model provides a close match to the most robust observational constraints currently available. Additionally, we find that: (a) SMBHs grow primarily through bursts of super-critical accretion ($\dot{M}_{\rm BH} > \dot{M}_{\rm Edd}$); (b) these bursts must be sufficiently long-lived, making the coherence timescale of accretion a key parameter in shaping the SMBH mass distribution; (c) the predicted $M_{\rm BH}$–$M_{\rm halo}$ relation remains approximately constant with redshift, with relatively tight scatter ($\lesssim 0.3$ dex) but extended high-SMBH-mass tails that give rise to luminous quasars; and (d) mergers contribute only marginally to SMBH growth compared to accretion, even under optimistic assumptions about merger timescales and remnant survival. The resulting SMBH growth histories, merger trees, and quasar light curves provide a versatile framework for future comparisons with the expanding range of observational constraints.

Work in preparation: **EP**, Joseph F Hennawi, Joop Schaye, et al., *Tracing individual black hole growth histories and quasar lightcurves in an N-body Universe*, Monthly Notices of the Royal Astronomical Society, *to be submitted.*
Reprinted here in its entirety.

5.1 Introduction

Understanding the growth and evolution of supermassive black holes (SMBHs) remains a central challenge in astrophysics. These objects are believed to reside at the centers of nearly all massive galaxies (e.g., Magorrian et al. 1998; Ferrarese & Merritt 2000; Kormendy & Ho 2013), where their accretion-powered emission gives rise to Active Galactic Nuclei (AGN) and quasars – some of the most luminous objects in the Universe (Salpeter 1964; Zel'dovich & Novikov 1967; Lynden-Bell 1969). Since the discovery of the first quasar (Schmidt 1963), a robust theoretical framework has emerged: the observed quasar radiation originates from the release of gravitational energy as matter accretes onto the black hole, with a small fraction of this rest-mass energy (known as *radiative efficiency*) converted into radiation, and the rest fueling SMBH growth.

Building on this theoretical framework, a foundational link between quasar activity and black hole growth was established by Soltan (1982). The Soltan argument posits that the redshift evolution of the quasar luminosity function (QLF) directly traces the accretion history of SMBHs. By integrating the observed quasar emission over cosmic time and assuming a value for the radiative efficiency, one can estimate the total black hole mass density accumulated during luminous accretion phases. This insight laid the foundation for a broader class of empirical models for SMBH evolution, which describe the growth of the black hole mass function by solving a continuity equation. These models constrain key physical parameters – such as radiative efficiency, duty cycle, and the Eddington ratio distribution – by requiring consistency between the observed quasar population across cosmic time and the local census of dormant black holes (e.g., Yu & Tremaine 2002; Merloni & Heinz 2008; Shankar et al. 2009; Aversa et al. 2015; Tucci & Volonteri 2017).

A complementary class of empirical models extends the treatment of SMBH evolution by explicitly linking it to galaxy formation, typically by assuming a redshift-dependent relationship between black hole and galaxy properties. These models build on well-established empirical connections between dark matter halos, galaxies, and SMBHs (e.g., Kormendy & Ho 2013; Reines & Volonteri 2015), and often adopt semi-empirical galaxy–halo frameworks (e.g., Behroozi et al. 2013) as a foundation. They then incorporate SMBH growth and quasar activity in a way that is consistent with both galaxy and black hole observables (e.g., Croton 2009; Conroy & White 2013; Caplar et al. 2015). A notable recent example is the TRINITY model (Zhang et al. 2023b), which jointly evolves the average properties of halos, galaxies, and SMBHs within a unified, data-constrained framework. By reproducing a wide range of observables – including galaxy stellar mass functions, quasar luminosity functions, and black hole–galaxy scaling relations – these models

provide powerful tools to investigate the co-evolution of halos, galaxies, and black holes across cosmic time.

The rationale behind empirical models stems from the recognition that a first-principles treatment of SMBH accretion in a cosmological context remains fundamentally out of reach. While supermassive black holes are routinely included in both semi-analytic models (SAMs) and cosmological hydrodynamical simulations – and AGN feedback is widely acknowledged as a key driver of galaxy evolution (e.g., Somerville & Davé 2015) – current theoretical frameworks are still limited in their ability to model SMBH growth in a self-consistent, physically grounded way. The core difficulty lies in the extreme dynamic range involved: black hole accretion unfolds on scales much smaller than a parsec (pc), yet it is regulated by – and feeds back into – processes acting on kiloparsec to tens-of-megaparsec scales. Bridging these disparate scales in a physically accurate manner remains computationally unfeasible, complicating the development of robust sub-grid (physical) models that can be reliably implemented in large-scale simulations and SAMs.

As a result, while sub-grid prescriptions for star formation processes have reached a relatively mature and consistent formulation, the modeling of black hole seeding, accretion, and feedback remains coarse and exhibits substantial variation across simulation platforms (Habouzit et al. 2021). It is therefore not surprising that even state-of-the-art cosmological simulations – despite their success in reproducing galaxy populations across cosmic epochs (Vogelsberger et al. 2020) – continue to struggle with matching SMBH and AGN observables. Most are calibrated to reproduce local SMBH-galaxy scaling relations (e.g., Di Matteo et al. 2005; Booth & Schaye 2009), yet yield widely divergent predictions for black hole growth and evolution at earlier times (Habouzit et al. 2021, 2022; Porras-Valverde et al. 2025).

To address these shortcomings, significant effort in recent years has gone into refining the treatment of AGN in cosmological simulations, with the goal of more accurately connecting black hole growth to the broader context of galaxy formation and evolution. Advances have been made across multiple fronts, including improved models for black hole seeding (e.g., Bhowmick et al. 2024), more accurate prescriptions for black hole dynamics (e.g., Genina et al. 2024), revised accretion models that go beyond the classical Bondi-Hoyle approach (e.g., Koudmani et al. 2024; Weinberger et al. 2025), and increasingly sophisticated implementations of AGN feedback (e.g., Huško et al. 2024). Nevertheless, fundamental uncertainties remain – particularly in the high-redshift regime – underscoring the continued need for flexible, data-driven models. In this landscape, empirical approaches serve as a valuable counterpart to simulations, extracting physical constraints directly from observations and offering an efficient means to explore parameter space. Ultimately, the combination of improved simulations and empirically

anchored models offers a promising path toward unraveling the complex history of SMBH growth across cosmic time.

In parallel with theoretical progress, the observational frontier has been advancing rapidly – particularly at high redshifts. Over the past few decades, large-scale spectroscopic surveys have revealed luminous quasars powered by SMBHs with masses $\gtrsim 10^9\,\mathrm{M}_\odot$ at $z \gtrsim 6$, during the epoch of reionization (Fan et al. 2006; Mazzucchelli et al. 2017b; Farina et al. 2022; Fan et al. 2023), and even out to $z \approx 7.5$, just 700 million years after the Big Bang (Bañados et al. 2018; Yang et al. 2020; Wang et al. 2021). These discoveries pose a significant challenge to conventional models of SMBH formation and growth. If black hole seeds originate from Population III stellar remnants with initial masses of $\sim 100\,\mathrm{M}_\odot$ (e.g., Heger et al. 2003), and accrete at the Eddington limit, there is simply not enough time to reach $\gtrsim 10^9\,\mathrm{M}_\odot$ by $z \sim 7$.

This tension has motivated the development of alternative scenarios for early black hole growth. Proposed pathways include the formation of massive seeds through the direct collapse of pristine gas clouds (e.g., Bromm & Loeb 2003; Volonteri et al. 2008; Latif & Ferrara 2016), runaway stellar mergers in dense nuclear star clusters (e.g., Omukai et al. 2008; Devecchi & Volonteri 2009), and sustained phases of super-Eddington accretion (e.g., Volonteri et al. 2015; Lupi et al. 2016; Inayoshi et al. 2016). Despite their theoretical appeal, however, current observations are insufficient to distinguish between these competing models. The problem is fundamentally degenerate: widely different combinations of initial seed masses, accretion rates, duty cycles, and merger histories can be fine-tuned to reproduce the observed SMBH population. Disentangling these scenarios and uncovering the early growth history of SMBHs requires new and more detailed observations.

Interestingly, new observational probes are now beginning to emerge. With the advent of JWST, AGN candidates hosting moderately massive black holes ($\sim 10^6$–$10^8\,\mathrm{M}_\odot$) have been identified at redshifts as high as $z \approx 8$–10 (e.g., Maiolino et al. 2024; Kokorev et al. 2023; Larson et al. 2023; Bogdán et al. 2024). Although the physical nature of many of these sources remains uncertain, ongoing and upcoming wide-field surveys with Euclid and the Roman Space Telescope are expected to deliver statistically robust samples of luminous quasars at the highest redshifts (e.g., Yang et al. in prep.). At the same time, different quasar observables beyond SMBH mass estimates – such as the luminosity function (Schindler et al. 2023; Matsuoka et al. 2023), Eddington ratio distributions (Wu et al. 2022), clustering (Arita et al. 2023; Eilers et al. 2024), and proximity zone sizes (Eilers et al. 2017, 2020; Ďurovčíková et al. 2024) – are now being extended to earlier epochs.

Emerging trends from these early data are already beginning to challenge traditional views of SMBH growth in the high-z Universe. In particular, recent clustering measurements indicate that luminous quasars at $z \approx 6-7$ exhibit surprisingly low duty cycles (Eilers et al. 2024; Pizzati et al. 2024b;

Huang et al. in prep.), suggesting that only a small fraction of SMBHs are actively accreting at any given time. These findings are consistent with independent constraints on quasar lifetimes derived from proximity zone sizes and damping wings in quasar spectra (Davies et al. 2019; Ďurovčíková et al. 2024), as well as from the spatial extent of the Lyα nebulae powered by quasar radiation (Ďurovčíková et al. 2025). Taken together, these results are difficult to reconcile with simple scenarios of continuous, Eddington-limited growth. Instead, they point toward a more nuanced evolutionary picture, in which SMBH accretion is stochastic and episodic, luminous phases are short-lived, and multiple growth pathways – including obscured or radiatively inefficient accretion – may contribute to the assembly of the most massive black holes in the early Universe.

At the same time, the analyses presented in Pizzati et al. (2024a,b, Chapters 2-3) embed the quasar population within a cosmological context, leveraging luminosity functions and clustering measurements at different redshifts – including the latest constraints at $z \gtrsim 6$. Using a consistent and homogeneous empirical framework, these studies uncover a puzzling trend: quasar properties, particularly their clustering, appear to evolve rapidly with redshift. This behavior is largely driven by the exceptionally strong clustering signal reported by Shen et al. (2007) at $z \approx 4$ using Sloan Digital Sky Survey (SDSS) data – a result that still awaits independent confirmation, but it nonetheless underscores our incomplete understanding of quasar evolution beyond cosmic noon. A key limitation of the Pizzati et al. (2024a,b) studies, however, is that each redshift is modeled in isolation, without tracing any evolutionary connection across epochs. Constructing a coherent, time-resolved model of SMBH growth is essential for interpreting these redshift trends within a unified framework.

Motivated by these considerations – and by the growing body of high-redshift observational constraints – we introduce a new empirical framework for modeling the evolution of SMBHs and quasars. Our model, BAQARO (Black Hole Accretion and Quasar Activity in a Realistic Observational framework), is designed to capture the early buildup of SMBH mass and the emergence of luminous quasars from cosmic dawn to cosmic noon. Quasar activity and SMBH growth are treated self-consistently, incorporating constraints not only from quasar luminosity functions but also from clustering measurements as well as from the distribution of SMBH masses and Eddington ratios at different epochs. This integrated approach allows us to explore a wide range of physical growth scenarios within a unified, observationally anchored framework.

The model is built on the merger trees and halo catalogs from the dark-matter-only (DMO) version of the FLAMINGO cosmological simulation suite (Schaye et al. 2023; Kugel et al. 2023). Specifically, we use the $(2.8\,\mathrm{cGpc})^3$ volume run, which offers the statistical power needed to sample the rare, luminous ($L_{\mathrm{bol}} \gtrsim 10^{47}\,\mathrm{erg\,s^{-1}}$) quasar population out to the highest redshifts.

By leveraging the output of large N-body simulations, we are able to trace SMBH accretion histories and quasar light curves along individual halo growth trajectories, naturally capturing the diversity and stochasticity of black hole evolutionary pathways. This is essential to recover the most massive SMBHs ($\gtrsim 10^9\,\mathrm{M}_\odot$) observed at early times, which are extreme outliers in the distribution of SMBH growth histories. Furthermore, modeling quasar clustering directly from the simulated large-scale structure eliminates the need for linear bias or halo model prescriptions, which are known to perform poorly for the halo mass and redshift regimes relevant to bright quasars (e.g., Mead & Verde 2021).

The paper is structured as follows. In Section 5.2, we introduce the key components of the BAQARO model and the data it aims to reproduce. In Section 5.3, we present the main results of our analysis, comparing our fiducial model with all observational constraints. Section 5.3.3 studies the implications of our model for the growth of SMBHs at early cosmic times and for the scaling relations between quasar/SMBH and halo properties. We summarize our findings and discuss our results from a broader perspective in Section 5.4.

5.2 Methods

At its core, BAQARO combines DMO cosmological simulations with a phenomenological prescription for black hole seeding and growth. We use the merger trees extracted from the FLAMINGO large-volume simulation to trace the assembly histories of dark matter halos, within which we model the evolution of SMBHs. Black holes are seeded in early halos and subsequently grow through a combination of gas accretion and black hole mergers. The accretion rate, together with an assumed radiative efficiency, determines the bolometric luminosity of each SMBH, allowing us to predict quasar light curves along with individual merger histories.

Our primary goal is to reproduce the statistical properties of the bright quasar population from cosmic dawn to cosmic noon. To this end, we calibrate our model to match three key observables across redshift: (i) the quasar luminosity function (QLF), (ii) the conditional Eddington ratio distribution function (cERDF), and (iii) the large-scale clustering of quasars. These constraints jointly inform the underlying growth histories of SMBHs and the physical conditions that shape quasar activity.

In the remainder of this section, we describe each component of the model in detail, beginning with the underlying simulation and halo merger trees, and proceeding through the prescriptions for seeding, merger, accretion, and quasar luminosity modeling.

5.2.1 Extracting halo mass histories and merger trees from the FLAMINGO simulation

To model the evolution of SMBHs and quasars in a cosmological context, we require a realistic description of the growth and assembly histories of dark matter halos across cosmic time. This is crucial because quasars are rare, highly biased tracers that inhabit the most massive and rapidly evolving structures in the Universe – environments whose complexity and stochasticity cannot be fully captured by analytic models of merger trees or average mass accretion histories (e.g., extended Press–Schechter; Lacey & Cole 1993). In particular, analytic approaches struggle to reproduce the nonlinear structure formation, mergers, and diverse growth trajectories of massive halos, especially at high redshift. Large cosmological N-body simulations, by contrast, can track halo growth and merger histories in detail, providing the physically grounded framework needed to model the evolution of luminous quasars in the Universe.

We obtain these halo growth histories from the DMO version of the FLAMINGO cosmological simulations (Schaye et al. 2023; Kugel et al. 2023), which combine the resolution and large volume necessary to capture the environments in which luminous quasars form and evolve. FLAMINGO is a suite of state-of-the-art simulations run with the SWIFT code (Schaller et al. 2024), which couples an N-body gravity solver with smooth particle hydrodynamics (SPH). Gravitational interactions are computed using the Fast Multipole Method (Greengard & Rokhlin 1987). The simulations adopt the "3×2pt + all" cosmological parameters from Abbott et al. (2022): $\Omega_\mathrm{m} = 0.306$, $\Omega_\mathrm{b} = 0.0486$, $\sigma_8 = 0.807$, $H_0 = 68.1\,\mathrm{km\,s^{-1}, Mpc^{-1}}$, and $n_\mathrm{s} = 0.967$, with a total neutrino mass of 0.06 eV. Massive neutrinos are included via the δf method of Elbers et al. (2021). Initial conditions are generated with multi-fluid third-order Lagrangian perturbation theory (3LPT), using partially fixed phases to reduce cosmic variance (Angulo & Pontzen 2016): the amplitudes of modes with $(kL)^2 < 1025$ are set to match the mean variance, where k is the wavenumber and L the box size.

In this work, we employ the DMO FLAMINGO run with a comoving box size of $L = 2.8\,\mathrm{cGpc}$, comprising 5040^3 cold dark matter (CDM) particles and 2800^3 neutrino particles. This corresponds to a CDM particle mass of $M_\mathrm{CDM} = 6.72 \times 10^9\,\mathrm{M_\odot}$, which – although relatively low in resolution – is sufficient to resolve the massive haloes expected to host luminous quasars. In future developments of our model, we plan to address this limitation by exploiting the newly developed FLAMINGO-10k simulation (Schaller et al., in prep.; Pizzati et al. 2024b), which contains eight times more particles and will enable us to trace SMBH growth starting from significantly lower-mass progenitors.

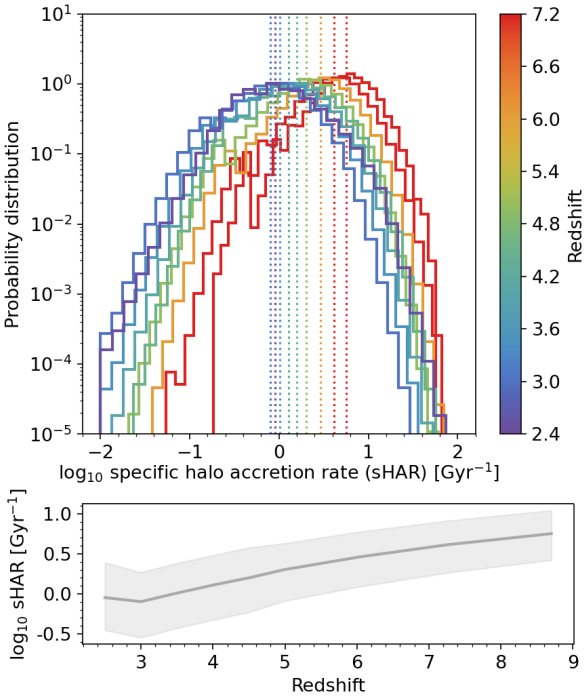

Figure 5.1: *Top*: Probability distribution of the specific halo accretion rate (sHAR) measured between two consecutive snapshots (Eq. 5.1). The corresponding redshifts are shown with different colors. The median values of the distributions are highlighted with dotted vertical lines. *Bottom*: Evolution of the median (solid line) and 16th-84th percentiles (shaded region) of the specific halo accretion rate as a function of redshift. The median specific halo accretion rate decreases by almost one order of magnitude between $z \approx 8.5$ and $z \approx 2.5$.

5.2.1.1 Subhalo masses and specific halo accretion rates

Our first step is to construct a comprehensive halo catalog across all simulation snapshots of interest. Specifically, we consider the 39 snapshots spanning from $z = 15$ – the highest redshift available in the simulation – down to $z = 2$. This lower redshift limit, which currently bounds our model to cosmic noon, is chosen to reduce computational cost; it will be extended in future iterations of the model. The snapshots are approximately evenly spaced in time, with a mean spacing of $\sim 80\,\mathrm{Myr}$. The time intervals in the redshift range considered vary from a minimum of 40 Myr to a maximum of 180 Myr.

We use the FLAMINGO halo catalogs generated with the HBT-HERONS code (Forouhar Moreno et al. 2025), an upgraded implementation of the Hierarchical Bound-Tracing (HBT) algorithm originally developed by Han

et al. (2012, 2018). HBT-HERONS identifies subhalos and follows their evolution across time by accounting for key physical processes such as mergers, accretion, and tidal stripping. To achieve this, it tags particles associated with a subhalo at a given snapshot based on their dynamical history, and then propagates these associations forward in time. In later snapshots, particles that originated from the same progenitor subhalo are regrouped to identify descendant subhalo candidates. This approach ensures temporal consistency and enables accurate reconstruction of merger trees for substructures in simulations. Notably, the resulting merger trees exhibit significantly fewer pathological failures – such as mass swapping or unphysical transients – than those created with other tree-building algorithms (Chandro-Gómez et al. 2025). These features make HBT-HERONS particularly well suited to our study, as capturing the hierarchical evolution of SMBHs depends critically on the fidelity of halo merger trees.

We adopt the bound mass definition for measuring subhalo masses: for each subhalo, the mass is computed by summing the gravitationally bound particles. However, satellite subhalos are often subject to strong tidal stripping, which can significantly reduce their instantaneous bound mass and thereby obscure their past gravitational influence. For this reason, we use the peak bound mass, $M_{\rm peak}$, defined as the maximum bound mass a subhalo has reached over the course of its assembly history. Conveniently, HBT-HERONS provides this quantity for each object by default, and we adopt it as our fiducial subhalo mass definition – i.e., we take $M_{\rm h} \equiv M_{\rm peak}$. Using this convention, we construct subhalo mass histories by following each object from the snapshot where it first appears through to the final snapshot included in our model, currently set at $z = 2$.

Although HBT-HERONS is capable of identifying subhaloes down to a minimum of 20 particles, we apply a stricter resolution cut to ensure the robustness of our results. Specifically, we exclude any halo that never exceeds 40 bound particles at any point between its formation and $z = 2$. This threshold removes transient or poorly resolved subhaloes that could otherwise introduce noise or instability, while retaining the well-resolved halos that meaningfully contribute to the buildup of supermassive black holes and the quasar population.

In addition to tracking subhalo masses, we also measure halo accretion rates, defined as the mass growth of subhaloes per unit time. These are computed directly from the merger trees by taking the difference in halo mass between two consecutive snapshots and dividing by the time interval separating them. In FLAMINGO, the typical snapshot spacing is $\sim 80\,{\rm Myr}$ (see above), which corresponds to a substantial fraction of the dynamical time for the halos of interest at the redshifts we study. This snapshot cadence allows us to resolve halo growth on timescales that are well-matched to the key cosmological processes governing halo and galaxy evolution – and, ultimately, the long-term fueling of SMBHs.

While the time spacing between snapshots in FLAMINGO is approximately constant, it is not strictly uniform across all redshifts. As a result, the accretion rates we compute are averaged over different time intervals at different epochs. In this initial analysis, we do not correct for this variation. However, because these differences are modest and the overall cadence remains physically meaningful, we do not expect this to introduce significant biases in our results. Nonetheless, we plan to implement redshift-dependent correction factors in future iterations to more rigorously account for this effect.

We emphasize that the computed accretion rates reflect the total mass growth of halos, including both smooth accretion and mergers. We do not attempt to separate these contributions, as both cosmological inflows and merger events are expected to play important roles in triggering black hole growth and quasar activity. For our purposes – linking SMBH accretion to host halo evolution – this combined accretion measure provides a physically motivated and practical proxy.

The top panel of Fig. 5.1 shows the distribution of specific accretion rates, $s\dot{M}_{\rm acc}$, for all halos in the simulation across a range of redshifts – corresponding to a subset of all the available snapshots. The specific accretion rate between two consecutive snapshots i and $i+1$ is defined as:

$$s\dot{M}_{\rm acc,i} = \frac{M_{{\rm h},i+1} - M_{{\rm h},i}}{(t_{i+1} - t_i)\, M_{{\rm h},i}}, \tag{5.1}$$

where $M_{{\rm h},i}$ and t_i are the halo mass and cosmic time at snapshot i, respectively. This quantity captures the relative growth rate of halos and serves as the foundation for our SMBH accretion prescriptions (Sec. 5.2.2).

The bottom panel of Fig. 5.1 illustrates the redshift evolution of halo growth by showing the median and scatter (standard deviation) of the specific accretion rate distribution as a function of redshift. As expected, typical accretion rates are higher at earlier cosmic times, reflecting the accelerated pace of structure formation in the high-redshift Universe (e.g., McBride et al. 2009).

5.2.1.2 Construction of the merger tree catalogs

Thanks to the structure of HBT-HERONS, merger trees are naturally constructed by following the evolution of subhalos over time through the tracker particle method described above. To build our merger tree catalog, we extract descendant information for each subhalo flagged as "dead" by the halo finder. This status is assigned when a subhalo is either gravitationally disrupted or sinks toward the center of a larger halo, meeting the merging criterion defined in Forouhar Moreno et al. (2025).

For subhaloes that have sunk, we directly identify the halo they merge into – referred to as the sink halo in the code. For those that are disrupted but

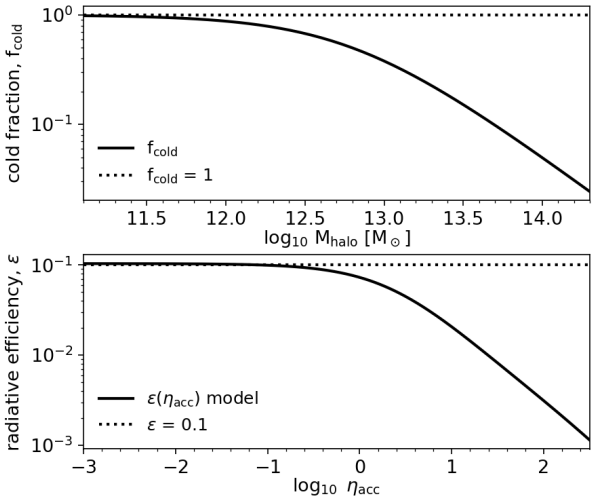

Figure 5.2: *Top*: Fraction of cold gas accreted onto halos, $f_{\rm cold}$, as a function of halo mass, $M_{\rm halo}$. The cold fraction accounts for the suppression of cold inflows in massive halos, where virial shock heating raises the gas temperature above the cooling threshold. We adopt the parametrization of Correa et al. (2018, see Eq. 5.5), assuming a redshift-independent $f_{\rm cold}$ since their model shows only weak evolution beyond cosmic noon. *Bottom*: Radiative efficiency, ϵ, as a function of the specific black hole accretion rate, $\eta_{\rm acc}$. We adopt the parametrization of Madau et al. (2014), fixing the spin parameter to $a = 0.67$, which yields a sub-Eddington efficiency of $\epsilon_0 \approx 0.1$. As $\eta_{\rm acc}$ approaches and exceeds the Eddington limit, ϵ decreases due to the transition to the slim-disk regime, where photon trapping and advective processes reduce the radiative output of the accretion flow (e.g., Sądowski et al. 2014).

not sunk, we search for a descendant subhalo that shares the majority of the tracker particles previously bound to the disrupted object. If such a match is found, it is designated as the descendant. If no suitable descendant can be identified based on particle overlap, we fall back to the HBT parent–child hierarchy, assuming the subhalo merges with its immediate parent as defined by the algorithm.

In a very small fraction of cases ($\lesssim 0.1\%$), no parent halo can be identified. These cases typically involve low-mass or field haloes that become tidally disrupted or were only transiently detected due to noise or numerical artifacts. We classify such objects as "lost", and we do not attempt to track the evolution of their associated SMBHs beyond the point of disruption.

5.2.2 Modeling SMBH and quasar evolution

Our model for the growth and radiative output of SMBHs is built around three key components: initialization, accretion (and the associated quasar

Table 5.1: Free parameters of the model, together with their values for the fiducial run analyzed in this work. A brief description of each parameter, along with the corresponding equation, is also provided.

Parameter	Value (fiducial model)	Description	Equation
$M_{\rm start}$	$5 \times 10^6\,{\rm M}_\odot$	SMBH mass initialized in newly-formed halos	-
$\tau_{\rm coherence}$	10 Myr	Coherence timescale of SMBH accretion	Eq. 5.12
$\eta_{\rm av,0}$	-1.12	Average sBHAR for $s\dot{M}_{\rm cold,acc} = 1\,{\rm Gyr}^{-1}$	Eq. 5.7
$\eta_{\rm av,evol}$	0.95	Power-law index of the average sBHAR – $s\dot{M}_{\rm cold,acc}$ relation	Eq. 5.7
σ_0	0.52	Scatter in the sBHAR for $s\dot{M}_{\rm cold,acc} = 1\,{\rm Gyr}^{-1}$	Eq. 5.8
$\sigma_{\rm evol}$	-0.17	Power-law index of the sBHAR scatter – $s\dot{M}_{\rm cold,acc}$ relation	Eq. 5.8

emission), and mergers. Each of these processes is detailed separately in the subsections that follow.

5.2.2.1 Black hole initialization

For the initialization of black holes, we adopt a simple prescription: a black hole is assigned to the center of each subhalo at the snapshot where the halo first appears in the merger tree. The black hole is initialized with a fixed mass, $M_{\rm start}$. This mass should not be interpreted as a physical seed mass in the early Universe, but rather as an empirical value that marks the beginning of the SMBH growth track within our model. A physical treatment of the seed mass regime would require either simulations with much higher mass resolution, capable of resolving halos down to $\sim 10^6 - 10^7\,{\rm M}_\odot$, or an analytical framework extending SMBH growth histories down to $10^2 - 10^4\,{\rm M}_\odot$.

Using a uniform value of $M_{\rm start}$ for all subhalos is, of course, a strong simplification. In reality, SMBH masses are expected to vary with host halo mass – which is similar for newly formed halos but not identical – as well as with the formation environment and underlying seeding channel (e.g., Li et al. 2021; Jeon et al. 2025). While intrinsic scatter in $M_{\rm start}$ could be easily incorporated into our framework to reflect these diverse formation pathways, its impact on our predictions is largely degenerate with scatter in the accretion rates. In practice, the quasar luminosities and final black hole masses are determined primarily by the integrated accretion history rather than the precise initial mass.

Nevertheless, the interplay between seeding and accretion remains an important open question for SMBH evolution models. In future work, we plan to explore alternative initialization prescriptions and to quantify the extent to which different assumptions about $M_{\rm start}$ can be disentangled from variations in accretion.

5.2.2.2 Black hole mergers

Black hole mergers are implemented in our model using a straightforward, mass-conserving prescription: when two host halos merge, we assume their central SMBHs merge instantaneously, and the remnant is assigned a mass equal to the sum of the progenitor masses. This simplifying assumption neglects the complex dynamical processes that, in reality, delay SMBH coalescence after the host halos (or galaxies) merge.

In hierarchical structure formation, the two SMBHs first sink toward the common centre via dynamical friction against the surrounding dark matter, gas, and stars (Begelman et al. 1980; Mayer et al. 2007). At kiloparsec to parsec scales, the pair forms a bound binary that hardens further through stellar scattering (e.g., Milosavljević & Merritt 2001) and/or interaction with circumbinary gas discs (e.g., Dotti et al. 2012). Only once gravitational

wave (GW) emission dominates the energy loss does the binary inspiral and coalesce. This multi-stage process can introduce delays ranging from a few hundred Myr to several Gyr between the halo merger and SMBH coalescence, depending on the host properties, gas content, and redshift.

The coalescence itself can impart a GW recoil velocity to the remnant SMBH due to the asymmetric emission of gravitational waves (Bekenstein 1973). Numerical relativity simulations show that these kicks can reach up to several thousand km/s for particular mass ratios and spin configurations (e.g., Herrmann et al. 2007). In massive galaxies, such velocities may cause the SMBH to oscillate about the galactic centre for hundreds of Myr, while in lower-mass systems they can exceed the escape speed, ejecting the SMBH entirely from its host (e.g., Blecha & Loeb 2008).

While these processes can be incorporated into physical models via phenomenological prescriptions (e.g., Volonteri & Rees 2006; Tanaka & Haiman 2009; Kelley et al. 2017), here we adopt the most conservative choice and neglect them. This effectively assumes zero delay between the subhalo merger and SMBH coalescence, and ignores the possibility of displacement or ejection. As such, our treatment yields an upper limit to the contribution of mergers to SMBH mass growth. As shown in Sec. 5.3.3.3, even under this optimistic assumption, mergers contribute only a subdominant fraction of SMBH growth across almost all redshifts of interest. The details of the merger prescriptions, therefore, have little influence on our main results, although they could become important in future work where we plan to study the SMBH merger rates and the occupation fraction of SMBHs in galaxy populations.

5.2.2.3 Black hole accretion and quasar radiation

In contrast to seeding and mergers, modeling black hole accretion – and the associated quasar activity – requires more detailed physical prescriptions. Accretion is the dominant growth channel for SMBHs over most of cosmic history and directly governs their observable luminosity output. Our approach is to connect the specific black hole accretion rate (sBHAR) to the specific halo accretion rate of the host subhalo, supplemented by a stochastic component that accounts for the scatter seen in observationally inferred accretion rates (e.g., Alexander et al. 2025).

We describe the probability distribution of the sBHAR as a conditional function, $P(\eta_{\rm acc} \,|\, s\dot{M}_{\rm cold,acc})$, where $\eta_{\rm acc}$ is the black hole accretion rate normalized to the Eddington rate, and $s\dot{M}_{\rm cold,acc}$ is the specific cold gas accretion rate onto the halo. The black hole specific accretion rate, $\eta_{\rm acc}$, is given by

$$\eta_{\rm acc} = \frac{\dot{M}_{\rm BH,acc}}{\dot{M}_{\rm Edd}}, \tag{5.2}$$

where $\dot{M}_{\rm BH,acc}$ is the SMBH mass accretion rate and $\dot{M}_{\rm Edd}$ is the Eddington accretion rate. The latter depends on the adopted radiative efficiency $\epsilon_0 = 0.1$ via $\dot{M}_{\rm Edd} = L_{\rm Edd}/\epsilon_0 c^2$, with the Eddington luminosity scaling linearly with black hole mass:

$$L_{\rm Edd} = \frac{4\pi G M_{\rm BH} m_p c}{\sigma_T} \approx 1.3 \times 10^{38} \left(\frac{M_{\rm BH}}{\rm M_\odot}\right) \,{\rm erg\,s^{-1}}, \tag{5.3}$$

where G is the gravitational constant, m_p is the proton mass, c is the speed of light, σ_T is the Thomson cross-section, and $M_{\rm BH}$ is the black hole mass.

The specific cold gas accretion rate onto halos, $s\dot{M}_{\rm cold,acc}$, is obtained by applying a cold fraction, $f_{\rm cold}$, to the total specific halo accretion rate, $s\dot{M}_{\rm acc}$ (Fig. 5.2, top panel):

$$s\dot{M}_{\rm cold,acc} = f_{\rm cold}\, s\dot{M}_{\rm acc}. \tag{5.4}$$

The cold fraction encapsulates the physical suppression of cold inflows in massive halos, where virial shock heating raises the gas temperature above the cooling threshold (Dekel & Birnboim 2006). In contrast, low-mass halos – particularly at high redshift – can sustain efficient cold gas accretion through narrow filaments of the cosmic web.

We model $f_{\rm cold}$ as a function of halo mass, $M_{\rm h}$, following the phenomenological fit proposed by Correa et al. (2018). Their model, calibrated on the EAGLE simulation (Schaye et al. 2015), reproduces the transition between cold-mode accretion in low-mass halos and hot-mode accretion in massive halos. While in general $f_{\rm cold}$ depends on both halo mass and redshift, the parametrization of Correa et al. (2018) shows only weak redshift evolution beyond cosmic noon. We therefore adopt a fixed functional form across cosmic time:

$$f_{\rm cold}(M_{\rm h}) = 1 - \frac{1}{1 + \left(\frac{M_{200}}{M_{1/2}}\right)^a}, \tag{5.5}$$

with $a = -1.07$ and $\log_{10}(M_{1/2}/{\rm M_\odot}) = 12.8$ fixed following Correa et al. (2018). In this model, $M_{1/2}$ marks the characteristic halo mass where half of the inflowing gas is in the cold phase, while a controls the steepness of the transition between cold and hot accretion regimes. As shown in the top panel of Fig. 5.2, the cold fraction declines sharply above $M_{\rm h} \sim 10^{12.5}\,{\rm M_\odot}$, reflecting the increasing dominance of virial shock heating in massive halos.

Because $s\dot{M}_{\rm cold,acc}$ is the primary driver of SMBH fueling in our framework, this scaling provides a direct link between large-scale halo growth and the small-scale accretion processes that power quasars. In this model, the black hole accretion rate depends solely on the host halo mass – through its cold gas fraction – and on its total accretion rate. We do not impose any explicit redshift dependence, based on the assumption that the fundamental physical mechanisms governing SMBH growth are not directly dictated by

cosmic time. Instead, redshift dependence emerges naturally through the evolution of the specific halo accretion rate itself (Fig. 5.1, bottom panel). As discussed in Sec. 5.3, this physically motivated, minimal prescription captures the key trends in SMBH and quasar evolution that our model is designed to reproduce.

The conditional accretion rate distribution, $P(\eta_{\rm acc}|s\dot{M}_{\rm cold,acc})$, can in principle take a variety of functional forms. We experimented with several parametrizations, including log-normal distributions, Schechter functions, and broken power laws. For the purposes of this work, we adopt a log-normal form, which we find to provide an adequate fit to the data while remaining mathematically simple. The conditional sBHAR distribution is thus written as

$$P(\log_{10}\eta_{\rm acc}|s\dot{M}_{\rm cold,acc}) = \\ = \frac{1}{\sqrt{2\pi}\sigma(s\dot{M}_{\rm cold,acc})}\exp\left(-\frac{\log_{10}^2\left(\eta_{\rm acc}/\eta_{\rm av}(s\dot{M}_{\rm cold,acc})\right)}{2\sigma^2(s\dot{M}_{\rm cold,acc})}\right) \tag{5.6}$$

where $\eta_{\rm av}$ and σ are the mean and log-normal scatter (expressed in dex) of the distribution, respectively.

In our implementation, the mean and scatter are each parametrized as a power-law function of the specific cold gas accretion rate:

$$\eta_{\rm av}(s\dot{M}_{\rm cold,acc}) = \eta_{\rm av,0}\left(\frac{s\dot{M}_{\rm cold,acc}}{1{\rm Gyr}^{-1}}\right)^{\eta_{\rm av,evol}} \tag{5.7}$$

$$\sigma(s\dot{M}_{\rm cold,acc}) = \sigma_0\left(\frac{s\dot{M}_{\rm cold,acc}}{1{\rm Gyr}^{-1}}\right)^{\sigma_{\rm evol}} \tag{5.8}$$

This choice keeps the model both flexible and interpretable: η_0 and σ_0 set the normalization of the distribution at a fiducial accretion rate of $1\,{\rm Gyr}^{-1}$, while $\eta_{\rm evol}$ and $\sigma_{\rm evol}$ control how the mean and scatter respond to changes in cold gas supply.

Although more elaborate functional forms are possible – including broken power laws or redshift-dependent terms – we find that this simple, four-parameter power-law scaling yields a satisfactory match to the observational constraints considered in this work. As shown later, it captures both the central trend and the dispersion of the sBHAR distribution across the relevant range of halo accretion rates.

Once the sBHAR distribution is specified, we can compute SMBH mass growth by accretion between two consecutive simulation snapshots. Assuming that $\eta_{\rm acc}$ remains constant between snapshots i and $i+1$, the black hole mass evolves according to the exponential growth expected for Eddington-limited accretion:

$$M_{\rm BH}(t_{i+1}) = M_{\rm BH}(t_i)\exp\left[\frac{t_{i+1}-t_i}{t_{\rm acc}(\eta_{\rm acc})}\right], \tag{5.9}$$

where $t_{\rm acc}(\eta_{\rm acc})$ is the Salpeter timescale corresponding to the chosen accretion rate.

The Salpeter timescale quantifies the e-folding time for black hole mass growth at a given $\eta_{\rm acc}$ and radiative efficiency ϵ:

$$t_{\rm acc}(\eta_{\rm acc}) = \frac{\epsilon}{(1-\epsilon)\eta_{\rm acc}} \frac{\sigma_T\, c}{4\pi G m_p} \approx \frac{4.5\times10^7\ {\rm yr}}{\eta_{\rm acc}} \left(\frac{\epsilon}{0.1}\right) \left(\frac{0.9}{1-\epsilon}\right), \quad (5.10)$$

The numerical approximation corresponds to the canonical Salpeter time for $\epsilon = 0.1$ and $\eta_{\rm acc} = 1$.

In general, the radiative efficiency ϵ is not a fixed quantity, but depends on the accretion rate, $\eta_{\rm acc}$. Both analytic arguments and numerical simulations indicate the existence of distinct accretion regimes with different radiative properties (e.g., Shakura & Sunyaev 1973; Abramowicz et al. 1988; Sądowski et al. 2014). At sub-Eddington rates ($\eta_{\rm acc} \lesssim 1$), accretion flows are typically radiatively efficient and well described by the geometrically thin, optically thick Shakura–Sunyaev disk model (Shakura & Sunyaev 1973). In this regime, ϵ is approximately constant – typically $\epsilon \sim 0.06$–0.3 depending on black hole spin – reflecting the high efficiency with which gravitational binding energy is converted into radiation (Thorne 1974).

As the accretion rate approaches and exceeds the Eddington limit ($\eta_{\rm acc} \gtrsim 1$), however, the efficiency can drop sharply. In this regime, accretion is often described by slim-disk models (Abramowicz et al. 1988; Sądowski et al. 2014) in which photon trapping becomes significant: radiation generated in the inner disk is advected inward with the gas rather than escaping. Combined with powerful radiation-driven outflows, this effect leads to a "saturated" luminosity that increases only logarithmically with $\eta_{\rm acc}$ (Ohsuga et al. 2005; Jiang et al. 2014). Consequently, the SMBH can experience rapid mass growth while radiating only modestly above the Eddington luminosity.

In this work, the adopted functional form of $\epsilon(\eta_{\rm acc})$ is shown in Fig. 5.2 (bottom panel). It reproduces the two main regimes outlined above: (i) a constant radiative efficiency at low and moderate accretion rates, in agreement with thin-disk theory, and (ii) a capped luminosity at super-Eddington rates, consistent with slim-disk prescriptions. We follow the parametrization of Madau et al. (2014), which empirically fits the results of general relativistic radiation-hydrodynamic simulations (Sądowski et al. 2014) as a function of black hole spin. Since our model does not track spin evolution, we adopt the curve corresponding to a fixed dimensionless spin parameter $a = 0.67$, yielding a constant sub-Eddington efficiency of $\epsilon_0 \approx 0.1$. This choice provides a smooth and physically motivated transition between the sub- and super-Eddington regimes, while preserving the correct asymptotic limits in both.

Because the time interval between consecutive simulation snapshots is typically much longer than the characteristic variability timescales of

SMBH accretion, assuming a fixed value of $\eta_{\rm acc}$ across an entire snapshot interval would be a poor approximation. Instead, we decide to "subcycle" the integration of each SMBH's mass history by introducing a shorter timescale, $\tau_{\rm coherence}$, which we interpret as the coherence timescale of the accretion process. Over each interval of length $\tau_{\rm coherence}$, the accretion rate is held constant; at the end of the interval, we draw a new, independent value of $\eta_{\rm acc}$ from the conditional sBHAR distribution.

In reality, SMBH accretion is a stochastic process spanning a wide hierarchy of variability timescales, from days to hundreds of million years. By adopting $\tau_{\rm coherence}$ as an effective coherence timescale, we approximate this stochasticity in a computationally tractable way, while preserving the statistical properties of the underlying accretion distribution. Further discussion on the role of $\tau_{\rm coherence}$ in the growth of SMBHs can be found in Sec. 5.3.3.2.

Between two consecutive simulation snapshots, i and $i+1$, we therefore draw

$$N = \left\lfloor \frac{t_{i+1} - t_i}{\tau_{\rm coherence}} \right\rfloor \tag{5.11}$$

independent values $\eta_{{\rm acc},j}$, and compute the SMBH mass at t_{i+1} as:

$$M_{\rm BH}(t_{i+1}) = M_{\rm BH}(t_i)\, \exp\left[\tau_{\rm coherence} \sum_{j=1}^{N} t_{\rm acc}^{-1}(\eta_{{\rm acc},j}) \right], \tag{5.12}$$

where $t_{\rm acc}(\eta_{{\rm acc},j})$ is the Salpeter timescale corresponding to the j-th sampled accretion rate.

At the same time, the sampled accretion rates determine the radiative output of quasars. In practice, we assign the bolometric luminosity of each SMBH at snapshot i using the most recent sampled accretion rate, $\eta_{{\rm acc},i}$, via:

$$L_{\rm bol}(t_i) = \epsilon(\eta_{{\rm acc},i})\, \eta_{{\rm acc},i}\, \dot{M}_{\rm Edd}\, c^2. \tag{5.13}$$

This framework yields, for each subhalo in the simulation, a self-consistent prediction for the SMBH mass, bolometric luminosity, and Eddington ratio at every snapshot. These can then be directly compared to observational constraints across cosmic time, enabling a joint test of both SMBH growth and quasar demographics in our model.

5.2.3 Overview of the observational constraints

We consider three key observational constraints, all targeting the luminous quasar population ($L_{\rm bol} \gtrsim 10^{45}\,{\rm erg\,s^{-1}}$): (i) the bolometric quasar luminosity function (QLF), (ii) the large-scale clustering of quasars, and (iii) the conditional Eddington ratio distribution function (cERDF) at fixed quasar luminosity.

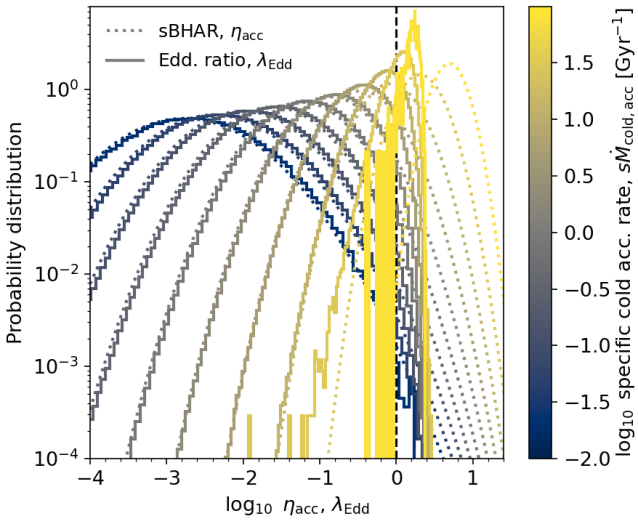

Figure 5.3: Probability distribution functions of the specific black hole accretion rate, $\eta_{\rm acc}$ (solid lines), and the Eddington ratio, $\lambda_{\rm Edd}$ (dotted lines), shown in bins of the specific cold accretion rate onto the halo. The $\eta_{\rm acc}$ distributions are drawn from the theoretical conditional probability $P(\eta_{\rm acc}|s\dot{M}_{\rm cold,acc})$ (Eq. 5.6), using the fiducial parameter values listed in Tab. 5.1. The $\lambda_{\rm Edd}$ distributions, in contrast, are measured directly from the model output. At sub-Eddington rates, they follow the same trend as $P(\eta_{\rm acc}|s\dot{M}_{\rm cold,acc})$ (with minor mismatches due to the non-uniform distribution of $s\dot{M}_{\rm cold,acc}$ within each bin), but they deviate markedly as $\lambda_{\rm Edd}$ approaches unity (dashed vertical line), reflecting the drop in radiative efficiency in this regime.

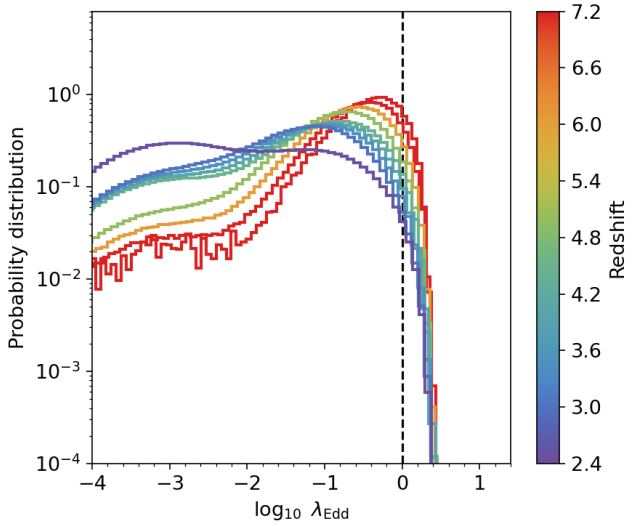

Figure 5.4: Probability distribution functions of the Eddington ratio, $\lambda_{\rm Edd}$, at different redshifts (solid colored lines) for the fiducial model described in Tab. 5.1. At high redshift, the majority of SMBHs accrete efficiently with $\lambda_{\rm Edd} \gtrsim 0.1$, whereas at later times accretion becomes progressively less efficient and the distribution shifts toward lower values. The dashed vertical line indicates the Eddington limit, $\lambda_{\rm Edd} = 1$.

Our first constraint is the bolometric QLF, which provides the most direct observable for comparison with theoretical models of quasar evolution. Unlike single-band surveys – particularly rest-frame UV selections, which systematically miss obscured quasars – the bolometric QLF combines multi-wavelength AGN datasets to reconstruct the full quasar population. Modern compilations draw on X-ray, mid-infrared, UV-optical, and radio observations (e.g., Hopkins et al. 2007a; Shen et al. 2020). X-ray data are especially valuable, as measurements of hydrogen column densities enable population-level obscuration corrections and yield intrinsic luminosities even for heavily absorbed sources (Ueda et al. 2014). By synthesizing these corrections across multiple bands, bolometric QLFs provide the most complete available census of SMBH accretion and thus a robust benchmark for testing models. In particular, using bolometric QLFs allows us to bypass the need for an explicit obscuration prescription in our framework. Likewise, the other two observables we consider – the clustering of quasars and the cERDF – remain unaffected by obscuration, provided that, at fixed luminosity, obscured and unobscured quasars represent a random subsample of the overall population.

Here, we adopt the bolometric QLF of Shen et al. (2020), which provides the most recent and comprehensive compilation of multi-wavelength data spanning $z \approx 0 - 6$. Since our focus is on luminous quasars ($L_{\rm bol} \gtrsim$

$10^{45}\,\mathrm{erg\,s^{-1}}$), the strongest constraints come from obscuration-corrected UV-optical samples such as those presented by Kulkarni et al. (2019), supplemented at high redshift by dedicated surveys (e.g., Matsuoka et al. 2018; Wang et al. 2019).

Although bolometric QLF reconstructions provide a substantially improved census of AGN activity and serve as a key benchmark for our model, significant uncertainties remain – particularly at high redshift, where both the fraction and physical nature of obscured quasars are still debated. In practice, X-ray-based obscuration corrections are possible only up to $z \lesssim 3$–4 and rely on uncertain extrapolations beyond this regime. Meanwhile, recent observations (e.g., Vito et al. 2018; Circosta et al. 2019; D'Amato et al. 2020; Gilli et al. 2022) and cosmological simulations (e.g., Ni et al. 2020; Vito et al. 2022; Bennett et al. 2024) suggest that quasars in the early Universe may be embedded in dense gas environments that drive high obscuration fractions, implying a rapid evolution of obscuration properties at $z \gtrsim 4$. If so, current bolometric QLF estimates at these redshifts likely underestimate the true space density of quasars, making them a conservative lower limit for comparison with theoretical models.

For the clustering of quasars, we rely on measurements of the two-point auto-correlation function from large spectroscopic surveys, which provide the most direct probe of quasar environments on cosmological scales. Numerous studies have characterized quasar clustering across a broad redshift range (e.g., Porciani et al. 2004; Croom et al. 2005; Porciani & Norberg 2006; Shen et al. 2007; da Ângela et al. 2008; Ross et al. 2009; White et al. 2012; Eftekharzadeh et al. 2015), consistently finding that luminous quasars typically reside in dark matter halos of mass $\sim 10^{12}$–$10^{13}\,\mathrm{M_\odot}$. In this work, which focuses on cosmic noon and earlier epochs, we adopt three key datasets: (i) the high-precision clustering constraints from the BOSS survey at cosmic noon ($z \approx 2.5$, Eftekharzadeh et al. 2015); (ii) the strong clustering of quasars at $z \approx 4$ reported by Shen et al. (2007) using SDSS measurements; and (iii) the recent quasar–galaxy cross-correlation measurements at $z \approx 6$ from the EIGER JWST survey (Eilers et al. 2024; Pizzati et al. 2024b). We do not include the auto-correlation function of faint quasars at $z \approx 6$ presented by Arita et al. (2023) because, as shown in Pizzati et al. (2024a, see their Appendix D), these data are not sufficiently constraining.

Each of the clustering measurements we consider applies a luminosity threshold, selecting quasars brighter than a given L_{bol}. Since quasar clustering may depend on luminosity, it is essential to adopt consistent thresholds when comparing our model predictions to the data. Accordingly, we impose the same luminosity cuts as in the observations, denoted $L_{\mathrm{bol,thr}}$. For the Eftekharzadeh et al. (2015) and Shen et al. (2007) measurements, we follow the thresholds used in the analysis of Pizzati et al. (2024a, see their Sec. 3.1): these are $\log_{10} L_{\mathrm{bol,thr}}/\mathrm{erg\,s^{-1}} = 46.1$ and $\log_{10} L_{\mathrm{bol,thr}}/\mathrm{erg\,s^{-1}} = 46.7$,

respectively. For the EIGER measurements of Eilers et al. (2024), we adopt the same threshold as in Pizzati et al. (2024b): $\log_{10} L_{\rm bol,thr}/{\rm erg\,s^{-1}} = 47.1$.

Finally, the cERDF is defined as the probability distribution of Eddington ratios at fixed bolometric luminosity, $P(\lambda_{\rm Edd} \,|\, L_{\rm bol})$. We use this quantity because it provides direct constraints on SMBH properties through broad-line measurements, while being robust to survey incompleteness. At fixed luminosity, the distribution of black hole mass estimates – and thus of $\lambda_{\rm Edd}$ – is determined primarily by the widths of broad emission lines in quasar spectra. Since these line widths are largely unaffected by survey flux limits, the cERDF is considerably less sensitive to selection effects than other diagnostics such as the total ERDF or the black hole mass function (BHMF).

We estimate the cEDRF using the SDSS quasar compilation in Wu & Shen (2022), which covers the redshift range $z \approx 0-6$. To extend the high-redshift coverage, we also incorporate the compilation of Fan et al. (2023), which includes all quasars known at $z > 5.9$ at the time of publication. In constructing the cERDF, we restrict the sample to sources with reliable SMBH mass estimates, and compute the Eddington ratio of each quasar as $\lambda_{\rm Edd} = L_{\rm bol}/L_{\rm Edd}$. Bolometric luminosities are derived from UV/optical magnitudes using the bolometric correction of Richards et al. (2006).

5.2.4 Fiducial model and parameter inference

The ultimate goal of our framework is to perform parameter inference and assess the predictive power of current quasar observables in constraining models of SMBH evolution. We plan to do so by writing a joint likelihood function for our model parameters, $\Theta : (M_{\rm start}, \tau_{\rm coherence}, \eta_{\rm av,0}, \eta_{\rm av,evol}, \sigma_0, \sigma_{\rm evol})$. The likelihood will incorporate the independent constraints coming from the three observables described in Sec. 5.2.3 – the QLF, the quasar auto-/cross-correlation functions ("corr"), and the cERDF:

$$\mathcal{L}^{(\rm total)} = \mathcal{L}^{(\rm QLF)}\,\mathcal{L}^{(\rm corr)}\,\mathcal{L}^{(\rm cERDF)} \tag{5.14}$$

The first two likelihood terms have the same expression:

$$\mathcal{L}^{(k)}(\mathbf{d}^{(k)}\,|\Theta) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})\right), \tag{5.15}$$

with $k \in \{\rm QLF, corr\}$, $\mathbf{d}^{(k)}$ being for the set of n data points with means $\mathbf{y}$ and covariance Σ coming from observations, and $\boldsymbol{\mu}$ the set of values predicted by our models.

For the third likelihood term, we directly compare the two-dimensional distribution in the bolometric luminosity-Eddington ratio plane predicted by our model with that measured from observations. Let $\mathbf{d}^{(\rm cERDF)} =$

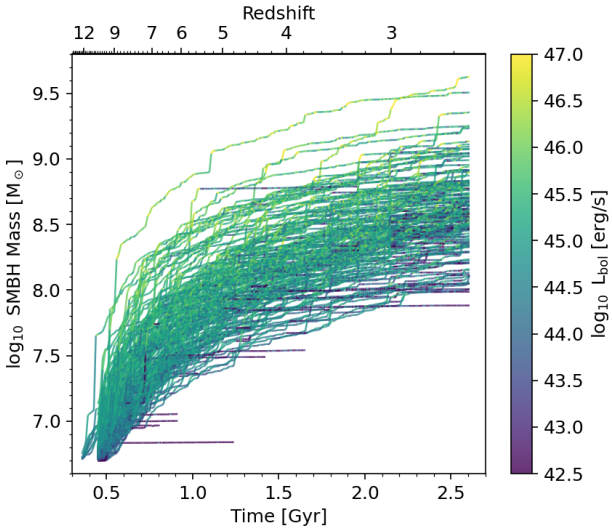

Figure 5.5: Evolution of black hole mass as a function of cosmic time (bottom axis) and redshift (top axis) for a subsample of 150 objects selected from the first 1000 halos to form in the simulation. Each track is color-coded by the instantaneous bolometric luminosity of the corresponding quasar. Tracks that terminate abruptly correspond to halos that either merge with more massive systems or are lost in the merger tree (see Sec. 5.2.1.2).

$\{(\lambda_{\mathrm{Edd},i}, L_{\mathrm{bol},i})\}$ denote the set of Eddington ratios and bolometric luminosities measured observationally, and let $P(\lambda_{\mathrm{Edd}}, L_{\mathrm{bol}})$ represent the corresponding probability distribution predicted by the model. The likelihood can then be written as:

$$\mathcal{L}^{(\mathrm{cERDF})}(\mathbf{d}^{(\mathrm{cERDF})}|\Theta) = \prod_i P(\lambda_{\mathrm{Edd},i}, L_{\mathrm{bol},i}) \tag{5.16}$$

While this inference roadmap is straightforward in principle, it is challenging to implement in practice. Performing inference requires evaluating the model many thousands of times across parameter space, which is computationally prohibitive: depending on the configuration, a single run of our model can take from several minutes to one hour. This makes direct inference with methods such as Markov Chain Montecarlo (MCMC) unfeasible.

To overcome this limitation, our long-term strategy is to construct a statistical emulator for the model observables, capable of reproducing the output of the full model at negligible computational cost. Such an emulator will enable efficient MCMC exploration of parameter space and a rigorous inference of posterior distributions. Development of this emulator is ongoing and will be included in future updates of this manuscript.

In the present study, we instead adopt a simpler approach: we fix the model parameters to fiducial values, chosen through a combination of trial-and-error exploration and preliminary experiments with the emulator. A systematic parameter inference study, together with an optimized parametrization of the accretion and merger processes, is deferred to forthcoming work.

The fiducial values used for all model parameters are listed in Table 5.1. The seed black hole mass is set to $M_{\rm start} = 5 \times 10^6\,{\rm M}_\odot$: given the resolution of the simulation, this corresponds to assigning each newly formed halo a black hole initially $\sim 5 \times 10^4$ times less massive than its host, broadly consistent with local SMBH–halo scaling relations. The coherence timescale of accretion, $\tau_{\rm coherence}$, is set to 10 Myr. This value ensures that individual high-accretion episodes are sufficiently long-lived to contribute meaningfully to SMBH growth, while remaining short compared to cosmological timescales. A more detailed discussion on $\tau_{\rm coherence}$ and our prior knowledge on the duration of SMBH accretion episodes is deferred to Sec. 5.3.3.2.

The ERDF parameters are set to the following fiducial values: $\eta_{\rm av,0} = -1.12, \eta_{\rm av,evol} = 0.95, \sigma_0 = 0.52, \sigma_{\rm evol} = -0.17$. This parametrization produces the conditional sBHAR distributions shown as dotted lines in Fig. 5.3. At high cold halo accretion rates ($\log_{10} s\dot{M}_{\rm cold,acc}/{\rm Gyr}^{-1} \gtrsim 0.5$), the sBHAR distribution becomes relatively narrow and centered at values close to or above the Eddington limit. As cold accretion onto halos declines, the distribution broadens and peaks at much lower sBHAR values.

The corresponding conditional distributions of the observed Eddington ratio, $\lambda_{\rm Edd} = L_{\rm bol}/L_{\rm Edd}$, are shown with solid lines. Importantly, $\lambda_{\rm Edd}$ coincides with $\eta_{\rm acc}$ only if the radiative efficiency is constant. In our model, where $\epsilon(\eta_{\rm acc})$ varies with accretion rate, $\lambda_{\rm Edd}$ no longer traces the intrinsic growth rate directly. This is evident in Fig. 5.3: while the $\lambda_{\rm Edd}$ distributions follow the sBHAR distributions at sub-Eddington rates (where ϵ is roughly constant), they remain sharply peaked near the Eddington limit even when the intrinsic sBHAR far exceeds unity. Physically, this reflects the saturation of radiative output in the super-Eddington regime, where SMBHs can accrete mass at highly efficient rates without producing proportionally higher luminosities. This effect has been invoked to explain the rapid assembly of massive black holes in the early Universe (e.g., Madau et al. 2014; Volonteri et al. 2015), and plays a central role also in our framework (see Sec. 5.3.1).

Figure 5.4 shows the distributions of observed Eddington ratios, binned by redshift. The model predicts a pronounced redshift evolution in SMBH accretion properties, driven primarily by the changing halo accretion rates over cosmic time (Fig. 5.1, bottom). At the epoch of reionization ($z \gtrsim 6$), the majority of SMBHs are actively accreting at $\lambda_{\rm Edd} \gtrsim 0.1$, consistent with rapid and efficient growth. By cosmic noon ($z \sim 2$), however, only a small fraction of SMBHs remain in the high-accretion regime, reflecting the global decline in gas accretion rates and SMBH activity. We note that very low

Eddington ratios ($\lambda_{\rm Edd} \lesssim 0.01$–$0.1$) are only useful for theoretical modeling as they are effectively unobservable in current quasar surveys. Consequently, the fraction of SMBHs accreting below this threshold defines an effective duty cycle of quasar activity in our framework. This duty cycle evolves rapidly with redshift, in agreement with expectations from measurements of the QLF and quasar clustering (e.g., Martini & Weinberg 2001; Haiman & Hui 2001; Pizzati et al. 2024b).

5.3 Results

We now turn to the results of the fiducial run introduced above. We begin by examining how SMBHs assemble their mass in our framework, focusing on the predicted accretion histories and comparing their radiative output to quasar observables. We then explore the broader implications of these results for the growth of SMBHs across cosmic history, with particular attention to the connection between black holes and the properties of their host halos.

5.3.1 The buildup of supermassive black holes across cosmic history

Figure 5.5 illustrates the mass assembly of SMBHs for a subsample of 150 objects, selected from the first 1000 halos formed in the simulation (i.e., halos that form at $z > 10$). The growth of these early black holes is initially rapid: within the first billion years, SMBH masses increase from the seed value $M_{\rm start}$ up to $M_{\rm BH} \approx 10^9\,{\rm M}_\odot$ by $z \approx 6$. This phase of accelerated growth coincides with sustained episodes of luminous quasar activity, with bolometric luminosities reaching $L_{\rm bol} \gtrsim 10^{47}\,{\rm erg\,s^{-1}}$. At lower redshifts, however, the buildup of SMBHs slows significantly. The vigorous accretion that characterizes the high-z regime – and enables the emergence of the first bright quasars – gradually gives way to a more quiescent growth pattern. By cosmic noon, quasars are powered less by rapid accretion and more by the sheer mass of their SMBHs: even relatively modest accretion rates can produce large luminosities once $M_{\rm BH}$ exceeds 10^8–$10^9\,{\rm M}_\odot$.

Overall, the bulk of the early-forming population grows from $M_{\rm start}$ at cosmic dawn to $M_{\rm BH} \approx 10^8$–$10^9\,{\rm M}_\odot$ by $z \sim 2$. Only a small subset of outliers – i.e., SMBHs experiencing unusually efficient or repeated accretion episodes – are able to reach the extreme SMBH masses associated with the brightest quasars observed across cosmic time. Tracking these rare growth histories, rather than focusing solely on population averages, is therefore essential for understanding the origin of luminous quasars.

These trends can be examined in more detail by following individual SMBH and quasar luminosity histories, as shown in Fig. 5.6. Here we select six halos that formed in the second snapshot of the simulation ($z = 12.26$) and

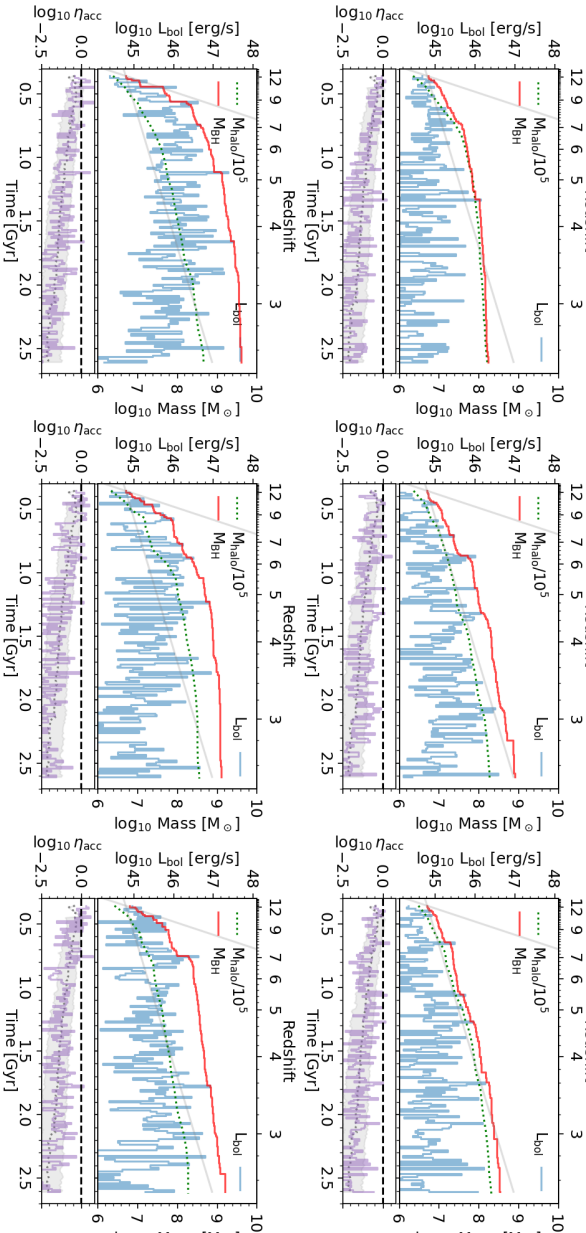

Figure 5.6: Individual accretion histories of SMBHs and their quasar light curves as a function of cosmic time (bottom axis) and redshift (top axis). We select six halos that form in the second snapshot of the simulation ($z = 12.26$). The SMBH mass is shown in red, alongside the host halo mass scaled down by 10^5 (green dotted). For reference, the steep and shallow gray lines indicate idealized SMBH growth tracks at constant accretion rates of $\eta_{\rm acc} = 1$ (Eddington rate) and $\eta_{\rm acc} = 0.1$, respectively. The quasar bolometric luminosity is plotted in blue; the blue and red curves intersect when quasars radiate at the Eddington limit. The bottom panels show the evolution of the specific black hole accretion rate, $\eta_{\rm acc}$, for each object (purple), compared to the population median (dotted gray line) and 16th–84th percentile range (shaded gray region).

plot their SMBH mass evolution (red) alongside the corresponding host halo mass (green, scaled down by 10^5). The associated quasar luminosity histories are shown in blue. Red and blue curves intersect when quasars radiate at the Eddington limit, while blue lines above the red indicate super-Eddington radiative phases. For reference, the light gray curves represent idealized cases of continuous SMBH growth from $M_{\rm start}$ at the Eddington limit and at $0.1\,\dot{M}_{\rm Edd}$. The smaller panels further show the accretion histories in terms of the sBHAR $\eta_{\rm acc}$ (purple), compared against the population median and 16th-84th percentiles (gray dotted line and shaded region) across cosmic time.

Across all examples, the same qualitative pattern emerges: rapid SMBH growth at $z \gtrsim 6$, followed by a marked slowdown at later times. This behavior is driven directly by the evolving distribution of sBHARs. At early times, SMBHs accrete at rates close to the Eddington limit; by lower redshifts, the distribution broadens and shifts toward low values ($\eta_{\rm acc} \lesssim 1\%$), consistent with the decline in cosmic gas supply. In most cases, SMBH growth closely tracks that of the host halo (green lines). However, stochastic variations in the accretion rate lead to significant departures in some cases.

Crucially, short bursts of super-critical accretion ($\eta_{\rm acc} > 1$) play a decisive role in driving SMBHs to higher masses on short timescales. For instance, the bottom-left panel of Fig. 5.6 shows an SMBH reaching $\sim 10^9\,{\rm M}_\odot$ at $z \approx 6$ through repeated episodes of super-critical accretion. Importantly, because the radiative efficiency ϵ drops steeply in the super-critical regime (Fig.5.2, bottom), very high accretion rates do not translate into equally high radiative output. Even when $\eta_{\rm acc} \gg 1$, the corresponding Eddington ratio saturates only modestly above unity (Fig. 5.3). This feature allows SMBHs in our model to gain mass rapidly at high redshift through brief high-accretion bursts ($1 \lesssim \eta_{\rm acc} \lesssim 10$), while remaining consistent with the empirical fact that strongly super-Eddington quasars are not observed at any epoch. In Sec. 5.3.2, we quantify this comparison by confronting our predicted Eddington ratio distributions with observations.

Interestingly, even with such strong accretion bursts, SMBH growth never systematically exceeds the analytic Eddington-limited growth curve (light gray). This highlights a central result of our model: sustained, uninterrupted Eddington-limited accretion is not a viable growth pathway. Instead, SMBHs grow through a stochastic sequence of accretion episodes – alternating between super-critical bursts and long periods of less efficient, sub-Eddington accretion. This produces growth tracks that may cluster around, but rarely exceed, the simple Eddington-limited scenario, while still enabling a subset of black holes to assemble the extreme masses required to power luminous quasars at all epochs.

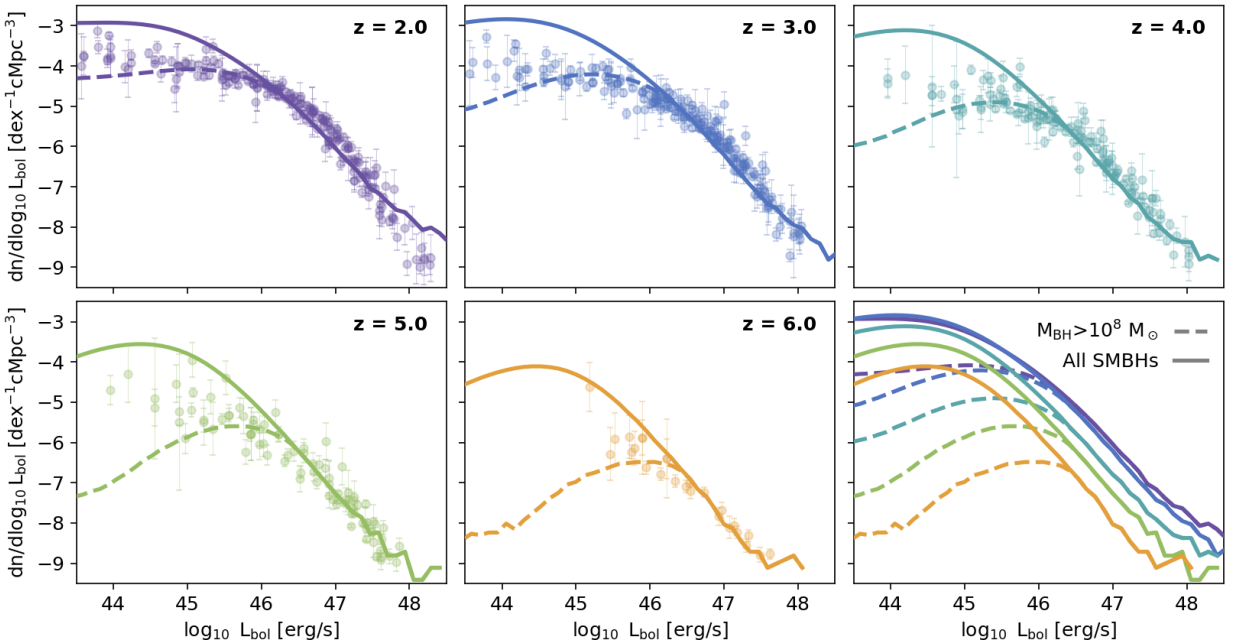

Figure 5.7: Comparison between the bolometric quasar luminosity function (QLF) predicted by our model and the compilation of Shen et al. (2020). Each panel corresponds to a different redshift, from $z = 2.0$ to $z = 6.0$, while the bottom-right panel combines all redshifts to illustrate evolutionary trends. Model predictions for the full SMBH population are shown as solid colored lines, and those restricted to $M_{\rm BH} > 10^8\,\mathrm{M}_\odot$ as dashed lines. Observational data – drawn from X-ray, UV–optical, infrared, and radio surveys, and corrected for obscuration using wavelength-dependent bolometric corrections – are shown as colored points with error bars.

5.3.2 Comparison with quasar observables

In Fig. 5.7, we compare the predictions of our fiducial model (solid lines) to the bolometric QLF of Shen et al. (2020, see Sec. 5.2.3). Each panel shows a different redshift slice from $z = 2$ to $z = 6$, while the bottom-right panel combines all redshifts to highlight the overall evolutionary trend.

We find that our model reproduces the observed QLF very well for luminous quasars ($L_{\rm bol} \gtrsim 10^{45.5}\,{\rm erg\,s^{-1}}$), which constitute the primary targets of UV-optical surveys. The large cosmological volume of FLAMINGO allows us to follow the quasar population up to extreme luminosities of $L_{\rm bol} \gtrsim 10^{48}\,{\rm erg\,s^{-1}}$, and down to space densities as low as $n \approx 10^{-9}\,{\rm cMpc^{-3}}$. This statistical power makes it possible to robustly probe the rarest, most luminous quasars, and we find that the bright-end slope of the QLF is accurately reproduced across all redshifts considered.

At the faint end, however, discrepancies arise. Observationally, the QLF flattens significantly toward low luminosities, whereas our model predicts a steeper continuation and thus an excess population of faint quasars/AGN, particularly at high redshift. This tension is not unique to our framework: many theoretical models have long struggled to reproduce the faint-end behavior of the QLF from first principles (e.g., Degraf et al. 2010).

It is important to emphasize that the faint end of the QLF is itself highly uncertain observationally. In UV-optical surveys, the QLF can be probed down to $L_{\rm bol} \approx 10^{45}\,{\rm erg\,s^{-1}}$, but at these luminosities quasars become increasingly difficult to distinguish from UV-bright galaxies, and completeness corrections are non-trivial. X-ray surveys extend coverage to lower luminosities, $L_{\rm bol} \sim 10^{43}$–$10^{44}\,{\rm erg\,s^{-1}}$, but rely heavily on uncertain photometric redshifts and rarely extend beyond $z \gtrsim 3$. Moreover, the level of obscuration at low luminosities and high redshifts is still poorly constrained: if the obscured fraction is higher than currently assumed (e.g., Ueda et al. 2014), present-day bolometric QLF estimates may underestimate the true space density of faint quasars.

The advent of JWST has opened a new observational window onto the faint end of the quasar population. Early studies identifying broad-line AGN through rest-frame optical diagnostics suggest the presence of a more numerous faint population than previously inferred, with bolometric luminosities of $L_{\rm bol} \sim 10^{43} - 10^{45}\,{\rm erg\,s^{-1}}$ and black hole masses of $M_{\rm BH} \sim 10^6 - 10^8\,{\rm M_\odot}$ (e.g., Harikane et al. 2023; Maiolino et al. 2024; Juodžbalis et al. 2025). Strikingly, our model naturally predicts such a population: low-mass SMBHs that were largely invisible to pre-JWST surveys make a substantial contribution to the faint end of the QLF. This is illustrated in Fig. 5.7 by comparing the full QLF (solid lines) with that obtained by restricting to SMBHs with $M_{\rm BH} > 10^8\,{\rm M_\odot}$ (dashed lines). The strong divergence of the two curves at $L_{\rm bol} \lesssim 10^{45}\,{\rm erg\,s^{-1}}$ highlights the predicted dominance of small SMBHs in this regime – consistent with the emerging JWST results.

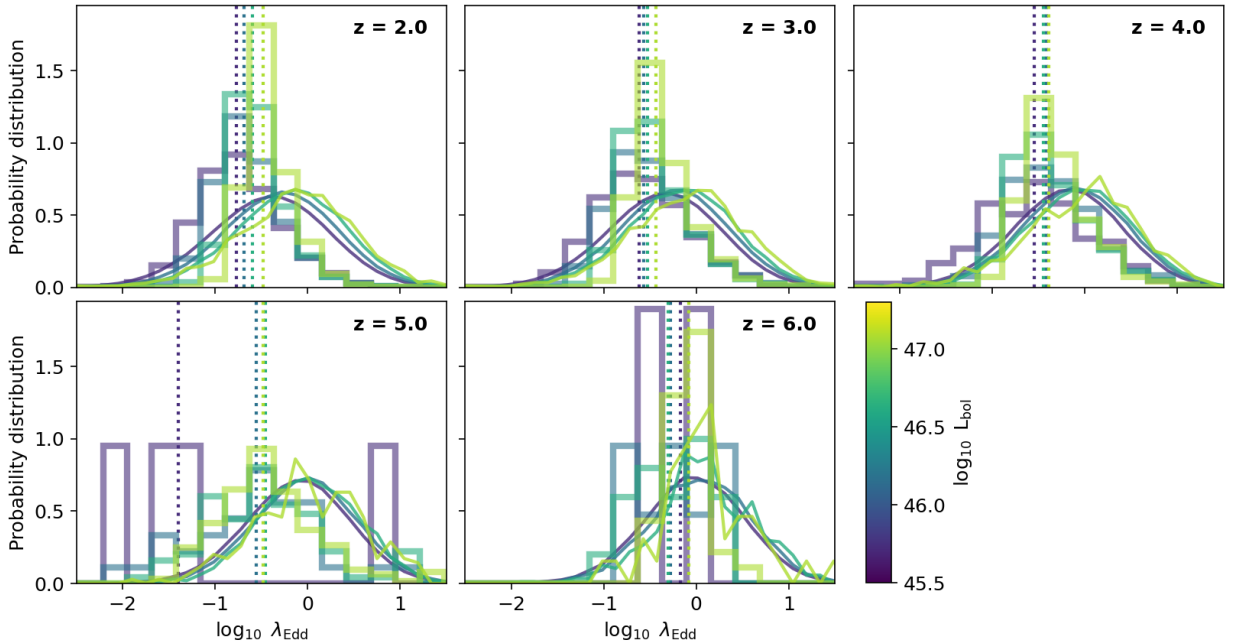

Figure 5.8: Comparison between the binned conditional Eddington ratio distribution functions (cERDFs) predicted by our model and those derived from SMBH mass and quasar luminosity compilations of Wu & Shen (2022, $z \approx 2$–5) and Fan et al. (2023, $z \gtrsim 6$). Observed distributions are shown as histograms, color-coded by bolometric luminosity, with dotted vertical lines marking their medians. Predictions from the fiducial model are shown as thin solid lines. While the model broadly reproduces the observed trends with both redshift and luminosity, it systematically predicts higher Eddington ratios than observed.

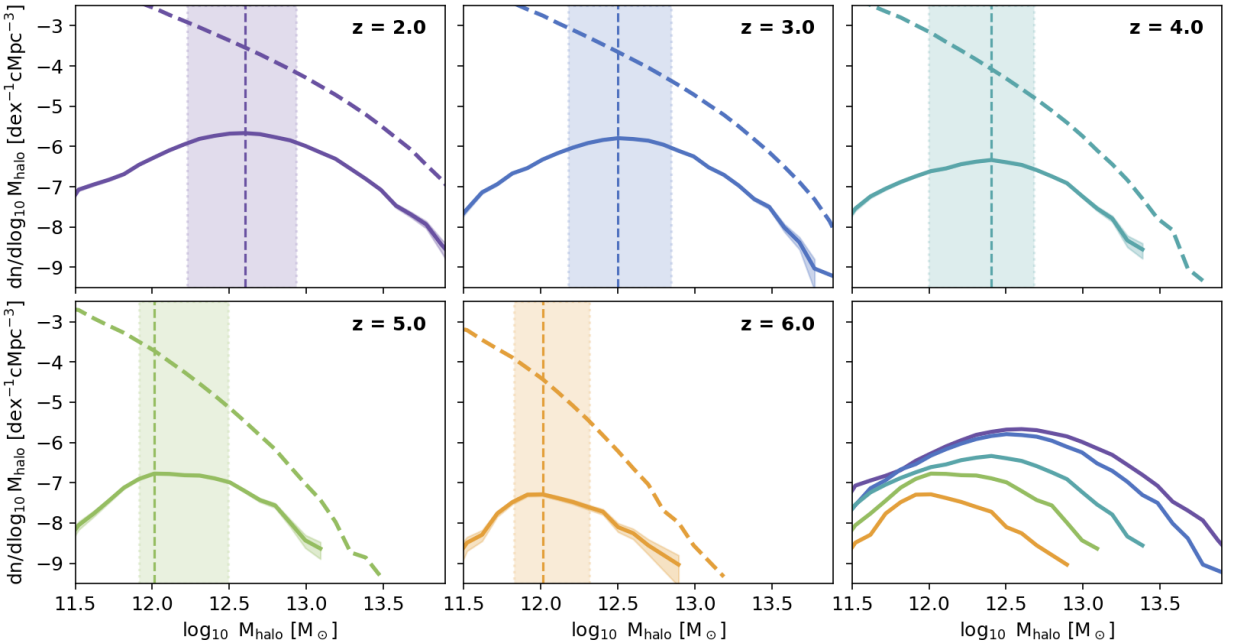

Figure 5.9: Quasar host mass functions (QHMFs) predicted by our fiducial model at different redshifts (solid colored lines). All QHMFs are computed using a uniform bolometric luminosity threshold of $\log_{10} L_{\rm bol,thr}/{\rm erg\,s^{-1}} = 46.5$. The bottom-right panel combines all redshifts to highlight evolutionary trends. For each distribution, vertical dashed lines mark the median, with shaded bands indicating the 16th–84th percentile range. The shaded envelopes around the solid lines represent Poisson uncertainties on the QHMF measurements. For comparison, the corresponding halo mass functions (HMFs) at each redshift are shown with dashed lines.

However, current JWST-based estimates remain highly uncertain, and the physical nature and demographics of these faint sources are still the subject of active debate.

If the faint QLF measurements in Fig. 5.7 are correct, then the faint-end excess predicted by our model reflects a genuine shortcoming of the framework. In this scenario, the model reproduces the QLF data across luminosities reasonably well only when restricted to $M_{\rm BH} > 10^8\,{\rm M}_\odot$ SMBHs (dashed lines). However, the additional population of lower-mass black holes accreting at or above the Eddington limit reaches luminosities comparable to faint quasars, thereby altering the predicted shape of the QLF. If such a population does not exist in reality, this would imply that SMBHs in our model are, on average, accreting – and thus radiating – more efficiently than is observed. This discrepancy could signal missing physical processes that regulate accretion in low-mass systems, such as stronger radiative or mechanical feedback, limited gas supply due to inefficient inflows, or environmental effects that suppress sustained high-Eddington accretion. More complex parametrizations are needed to account for these effects in the context of our framework.

A discrepancy between our model predictions for SMBH accretion and observations becomes indeed apparent when examining the cERDF in Fig. 5.8, where we show the distribution of Eddington ratios in bins of bolometric luminosity. In practice, the two-dimensional distribution $P(\lambda_{\rm Edd}, L_{\rm bol})$ is projected into a set of one-dimensional distributions by slicing along narrow $L_{\rm bol}$ bins. Comparing the observed distributions (histograms) with those predicted by the model (thin solid lines) reveals a systematic offset: the model consistently produces Eddington ratio distributions skewed toward higher values relative to the data. The difference is modest – the peaks of the two distributions typically agree within 1σ – but its persistence across all redshifts and luminosities suggests that it reflects a genuine limitation of the model rather than statistical noise.

This systematic bias toward higher $\lambda_{\rm Edd}$ is also what drives the large population of lower-mass SMBHs in our model to reach bolometric luminosities comparable to those of faint quasars. Reducing the average accretion rates would suppress the number of such faint sources and improve agreement with observations, but it would also prevent SMBHs from growing rapidly enough to reach the extreme masses required to power the brightest quasars. This underscores a fundamental challenge: reconciling the high accretion rates seemingly necessary to assemble billion-solar-mass black holes at early times with the empirical evidence that most observed quasars radiate at Eddington ratios around or below unity. Although our prescription for radiative efficiency in the super-critical regime (Madau et al. 2014) suppresses the luminosities of rapidly accreting quasars, limiting them to radiate only modestly above the Eddington limit (Fig. 5.3), we conclude that this effect alone does not fully reconcile the tension between modest observed Edding-

ton ratios and the growth rates required for SMBH assembly. Resolving this discrepancy will require further work to test whether refined parameter choices, alternative radiative efficiency prescriptions, or more flexible accretion models can provide a better match.

Despite this offset, our fiducial model successfully reproduces the main behavior of the cERDF. In particular, it captures the observed trend that the cERDF peaks at higher $\lambda_{\rm Edd}$ with both increasing redshift and increasing bolometric luminosity. This agreement suggests that while the model slightly overestimates accretion efficiencies, it nonetheless recovers the key qualitative features of SMBH growth across cosmic time and luminosity, making it a solid foundation for future refinements.

The last observable we consider is the clustering of quasars. As discussed in Sec. 5.2.3, we include measurements of the quasar auto-correlation function at $z \approx 2.5$ and $z \approx 4$, as well as the quasar–galaxy cross-correlation function at $z \approx 6$. In principle, the auto-correlation could be computed directly from the quasars in our model above a chosen luminosity threshold, $L_{\rm bol,thr}$. However, this approach is not feasible for the quasar-galaxy cross-correlation, since the FLAMINGO run used here does not resolve the majority of galaxy-hosting halos[1]. To ensure consistency across all redshifts, we instead adopt the framework developed by Pizzati et al. (2024a,b), and compute clustering predictions from the quasar host mass function (QHMF)[2].

The QHMF is defined as the halo mass distribution of quasars brighter than $L_{\rm bol,thr}$. Given a QHMF, the clustering can be predicted under the assumption that halo mass alone determines the bias – i.e., neglecting any assembly bias contributions (e.g., Wechsler et al. 2006). This assumption is reasonable in our context, as assembly bias is expected to play a minor role, and current quasar clustering measurements are not yet precise enough to be dominated by such effects (e.g., Bonoli et al. 2010). Using the halo correlation fitting framework of Pizzati et al. (2024a), we compute the quasar auto-correlation functions at $z \approx 2.5$ and $z \approx 4$ given the QHMFs at the respective redshifts (see their Eqs. 7–8). For the quasar–galaxy cross-correlation at $z \approx 6$, we combine the QHMF with a galaxy host mass function (GHMF) following Eqs. 3–5 in Pizzati et al. (2024b). For the GHMF, based on [O III] -emitting galaxies, we adopt the measurements of Eilers et al. (2024), and approximate it with a simple cutoff form in which the HMF is set to zero below $\log_{10} M_{\rm min,gal}/{\rm M}_\odot = 10.56$.

[1] By extending our framework to the larger FLAMINGO-10k simulation (Schaller et al., in prep.; Pizzati et al. 2024b), we will be able to compute the quasar-galaxy cross-correlation function directly from the simulation outputs. Work on this extension is currently underway.

[2] Pizzati et al. (2024a,b) also rely on the FLAMINGO suite of cosmological simulations. In their approach, the simulations are used to calibrate a fitting model that predicts the clustering of any halo population (both auto- and cross-correlations) directly from its mass distribution.

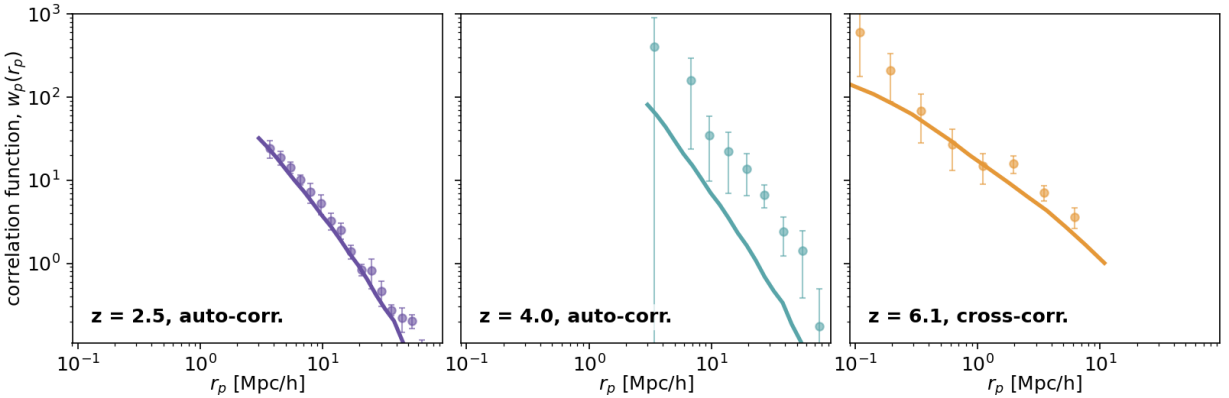

Figure 5.10: Comparison between quasar clustering predicted by our model and observational measurements at three redshifts: $z \approx 2.5$ (quasar auto-correlation function from Eftekharzadeh et al. 2015), $z \approx 4$ (quasar auto-correlation function from Shen et al. 2007), and $z \approx 6$ (quasar–galaxy cross-correlation function from Eilers et al. 2024). Model predictions are derived from the quasar host mass functions (QHMFs), adopting luminosity thresholds matched to each dataset, as described in Sec. 5.2.3.

Figure 5.9 shows the QHMF (solid lines) from our fiducial run, using a uniform luminosity threshold of $\log_{10} L_{\rm bol,thr}/{\rm erg\,s^{-1}} = 46.5$ at all redshifts. For comparison, the halo mass function (dashed line) is also shown, illustrating the fraction of halos active as quasars at a given mass. This fraction can be interpreted as the quasar duty cycle (e.g., Martini & Weinberg 2001; Haiman & Hui 2001; Pizzati et al. 2024a). At all redshifts, the QHMF peaks in the range $M_{\rm halo} \sim 10^{12}$–$10^{13}\,{\rm M}_\odot$, in good agreement with general observational trends. As redshift decreases, the QHMF peak shifts to progressively higher halo masses, and the distribution broadens, reflecting the growing diversity of quasar host environments.

Finally, Fig. 5.10 compares the clustering predicted by our model with the observational measurements described in Sec. 5.2.3. To ensure consistency, we compute clustering from the QHMFs using the same luminosity thresholds as the data at each redshift. This differs slightly from the QHMFs shown in Fig. 5.9 (that are obtained with a uniform luminosity threshold), but in practice the impact is modest: in our framework, quasar clustering depends only weakly on luminosity, in agreement with observations that find a mild or negligible luminosity dependence (e.g., Adelberger & Steidel 2005; Porciani et al. 2004; Shen et al. 2009; Eftekharzadeh et al. 2015).

At $z \approx 2.5$, our fiducial run reproduces very well the tight clustering constraints from the BOSS survey (Eftekharzadeh et al. 2015). Indeed, Pizzati et al. (2024a) derived the QHMF at this redshift by jointly fitting the QLF and clustering, finding a broad distribution peaking at $M_{\rm halo} \approx 10^{12.5}\,{\rm M}_\odot$, in excellent agreement with our model predictions.

At $z \approx 4$, by contrast, our model underpredicts the remarkably strong clustering measured by Shen et al. (2007) from SDSS quasars. Those data imply a very rapid evolution of quasar bias with redshift – a result that has long posed challenges for models of quasar and SMBH evolution (e.g., White et al. 2008; Shankar et al. 2010b). More recent measurements at comparable (He et al. 2018; Timlin et al. 2018) and higher redshifts (Eilers et al. 2024; Pizzati et al. 2024b) indicate weaker clustering, casting doubt on the extreme values implied by Shen et al. (2007). It is therefore not surprising that our model – like many other empirical models and SAMs (e.g., Conroy & White 2013; Fanidakis et al. 2013) – predicts significantly lower clustering at $z \approx 4$. Matching the Shen et al. (2007) result would require an extremely narrow QHMF, with virtually all luminous quasars confined to the most massive halos (Pizzati et al. 2024a). Such a scenario is difficult to reconcile with the intrinsic stochasticity that, in our model, drives extreme SMBH growth at early cosmic times (Sec. 5.3.3.1). The presence of stochasticity inevitably lowers the clustering, as it implies that even lower-mass halos can host massive, highly accreting SMBHs. This tension highlights the need for further work to establish whether the $z \approx 4$ clustering measurements can be reconciled with the broader set of quasar and SMBH constraints, and calls for new clustering analyses at similar redshifts

– such as those anticipated from DESI (Yang et al. 2023) – which will be essential to determine whether the strong $z \approx 4$ signal reflects genuine quasar physics or arises from observational systematics.

At $z \approx 6$, the agreement improves again. The EIGER survey (Eilers et al. 2024) measures a clustering signal somewhat stronger than our prediction, consistent with a QHMF peaking near $\approx 10^{12.5}\,\mathrm{M}_\odot$ (Pizzati et al. 2024b). Our model predicts a slightly lower peak halo mass and weaker clustering, but the discrepancy is modest, and a refined choice of model parameters would likely bring the results into even closer agreement. Moreover, the EIGER result is based on only five quasar fields and is therefore highly sensitive to cosmic variance. In future iterations, we will compare against forthcoming results from the JWST ASPIRE survey (Wang et al. 2023, Wang et al. in prep.), which will provide clustering measurements from a much larger sample of 25 quasars. These data, which are largely consistent with EIGER but significantly more robust against cosmic variance (Huang et al. in prep.), will offer a more stringent benchmark for constraining our model at early cosmic times.

5.3.3 Implications for SMBH growth and scaling relations

The central question we set out to address in this work is straightforward: can a simple, physically-motivated model for black hole formation and evolution reproduce the diverse properties of bright quasars observed across cosmic time? While additional work is required to perform a full inference analysis and refine the model parameters, the results presented in Sec. 5.3.2 demonstrate that the answer is encouragingly positive. Our framework successfully captures the key observational benchmarks – the quasar luminosity function, the conditional Eddington ratio distribution, and quasar clustering – for a wide redshift range.

Building on this result, we now turn to the broader implications of the model. In particular, we examine how our framework informs the scaling relations between SMBHs and their host halos, and what it reveals about the physical processes that govern SMBH growth across cosmic history. By connecting the global quasar population to the detailed assembly histories of halos, our model provides a natural way to probe both the average evolutionary pathways and the stochastic variability that drive the emergence of the most massive black holes.

5.3.3.1 The black hole mass-halo mass relation across cosmic history

Figure 5.11 shows the black hole mass–halo mass (M_{BH}–M_{halo}) relation predicted by our fiducial model. Each panel displays the full distribution of

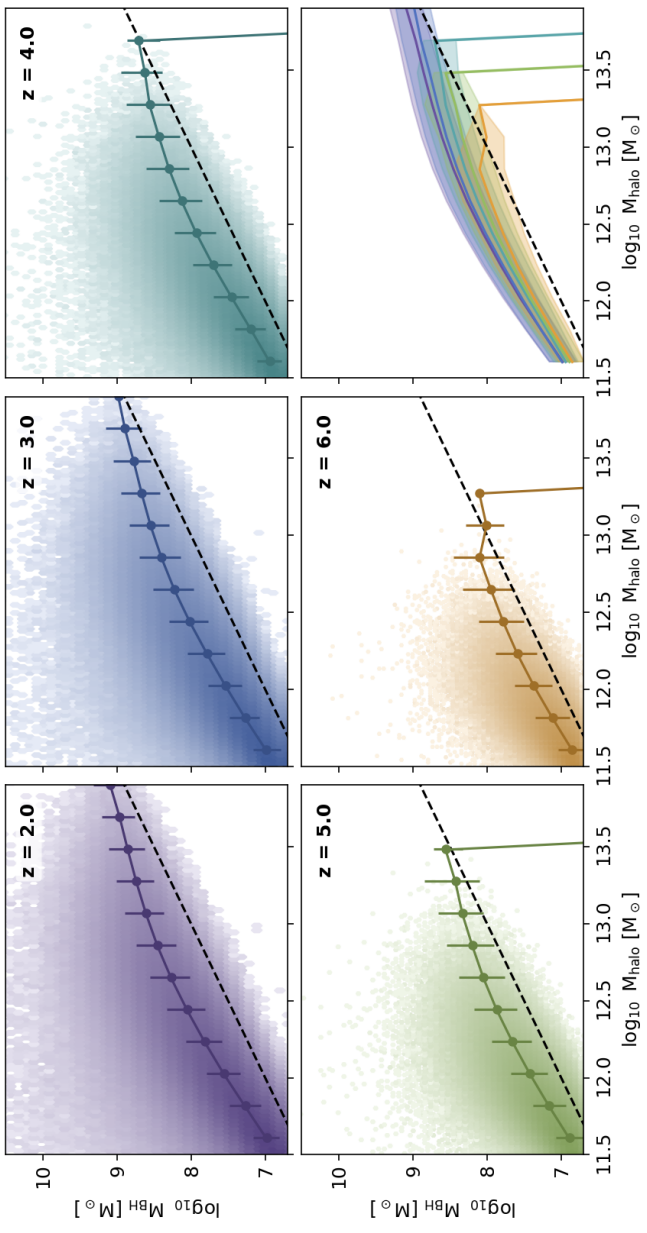

Figure 5.11: Black hole mass–halo mass relation predicted by our fiducial model at different redshifts. The two-dimensional distribution of SMBHs in the $M_{\rm BH}$–$M_{\rm halo}$ plane is shown with a logarithmic color scale. Median values of $M_{\rm BH}$ in bins of halo mass are plotted as points, with error bars indicating the 16th–84th percentile range. For reference, the dashed line in each panel marks a linear relation with normalization $M_{\rm BH}/M_{\rm halo} = 10^{-5}$. The bottom-right panel combines all redshifts, showing median relations (solid lines) and their scatter (shaded regions) to highlight evolutionary trends.

SMBHs at a given redshift as a two-dimensional histogram in the $M_{\rm BH}$–$M_{\rm halo}$ plane (with logarithmic color scaling to highlight the tails), while points and error bars denote the median and 16th–84th percentiles of $M_{\rm BH}$ in bins of halo mass. For reference, a linear relation with normalization $M_{\rm BH}/M_{\rm halo} = 10^{-5}$ is shown as a dashed line.

It is important to stress that the relation shown here cannot be directly compared to observations. In reality, SMBH mass measurements are subject to systematic uncertainties of order ~ 0.5 dex, and only a biased subset of the population is accessible – either luminous quasars radiating above survey thresholds or, in the nearby Universe, the most massive black holes detectable through dynamical methods. By contrast, our model includes the entire SMBH population, independent of observability, and assumes perfect knowledge of their masses.

With these caveats in mind, the relation exhibits a clear, nearly linear trend with relatively small scatter ($\lesssim 0.3$ dex). The very tight distribution at low SMBH and halo masses primarily reflects our assumption of a fixed seeding mass, $M_{\rm start}$. At larger masses the scatter increases moderately, but the relation remains well-defined, a direct consequence of our prescription that ties SMBH growth to halo accretion. This coupling is also evident in Fig. 5.6, where the growth of individual SMBHs broadly parallels the assembly of their host halos.

The bottom-right panel of Fig. 5.11 highlights the redshift evolution of the median relation and its scatter. Overall, the evolution is modest: at $M_{\rm halo} \lesssim 10^{13}\,{\rm M}_\odot$, both the normalization and slope of the relation increase toward lower redshifts. At the high-mass end, by contrast, the relation shows a clear flattening, which becomes more apparent once massive halos emerge in significant numbers near cosmic noon. This flattening is a direct consequence of our cold-gas accretion prescription: once halos exceed $\sim 10^{12.5-13}\,{\rm M}_\odot$, the smaller cold gas accretion rates limit the ability of SMBHs to grow in lockstep with their hosts. As a result, the most massive halos host SMBHs that grow more slowly relative to halo mass assembly. This behavior mirrors the turnover observed in the stellar-to-halo mass relation (e.g., Behroozi et al. 2019), and underscores a common physical picture in which cooling inefficiencies in massive halos suppress baryonic growth across both galaxies and SMBHs.

Despite the relatively small scatter, most SMBHs in our model remain below $M_{\rm BH} \sim 10^9, {\rm M}_\odot$ across all redshifts and halo mass bins, with the median relation at high redshift reaching only $M_{\rm BH} \sim 10^8, {\rm M}_\odot$ even in the most massive halos. The billion-solar-mass SMBHs powering luminous quasars at early times are therefore not typical products of the mean relation, but instead arise as stochastic outliers in the accretion history distribution. Indeed, the rare objects with $M_{\rm BH} \gtrsim 10^9\,{\rm M}_\odot$ are found in a wide range of halo masses, indicating that their rapid growth is driven more by fluctuations in accretion than by steady halo mass assembly. This finding reinforces the

importance of tracing individual SMBH growth trajectories – rather than relying solely on population averages – to capture the formation pathways of luminous quasars.

This intrinsic stochasticity also explains why our model struggles to reproduce the strong clustering signal measured at $z \approx 4$ by Shen et al. (2007, Sec. 5.3.2). Matching such strong clustering would require the massive SMBHs powering quasars to reside exclusively in the most massive halos – contrary to the broad range of environments predicted here. One could, in principle, reduce stochasticity by shifting the median relation upward (i.e., assuming more efficient accretion on average), or by allowing SMBHs in massive halos to continue accreting by relaxing the cold-gas suppression. However, both approaches would likely lead to an overproduction of extremely massive black holes at later times, in conflict with constraints in the local Universe. While low-redshift data are not explicitly included here, future work will explore whether low-z constraints, such as the local $M_{\rm BH}$–$M_{\rm halo}$ relation (e.g., Ferrarese & Merritt 2000), can help anchor the high-redshift regime and clarify which pathways of early SMBH growth remain consistent with observational constraints at later cosmic time.

5.3.3.2 The coherence timescale of the accretion process

A key driver of stochasticity in our SMBH growth model is the coherence timescale of the accretion process, $\tau_{\rm coherence}$. This often-overlooked parameter sets the degree of temporal correlation in SMBH accretion. While the intrinsic shape of the accretion rate (sBHAR) distribution, $P(\eta_{\rm acc}|s\dot{M}_{\rm cold,acc})$, specifies only the zeroth moment of the accretion stochastic process, $\tau_{\rm coherence}$ encodes its higher-order temporal structure, determining how fluctuations are sampled and accumulated over time.

If $\tau_{\rm coherence}$ is large, accretion bursts persist for extended periods; over the interval between two snapshots, the accretion history is then determined by only a few draws from the $P(\eta_{\rm acc}|s\dot{M}_{\rm cold,acc})$ distribution, yielding substantial object-to-object scatter in final SMBH masses. Conversely, a very small $\tau_{\rm coherence}$ yields many (approximately) independent draws in a fixed interval, so individual histories converge toward the mean behavior of the distribution. As Eq. 5.12 makes explicit, SMBH growth is governed by the sample mean of the $t_{\rm acc}^{-1}(\eta_{\rm acc})$ distribution. If $N \simeq \Delta t/\tau_{\rm coherence}$ is the effective number of independent draws over a snapshot interval Δt, then the standard deviation of the sample mean distribution scales as $N^{-1/2} \propto (\tau_{\rm coherence}/\Delta t)^{1/2}$ – directly linking larger $\tau_{\rm coherence}$ to larger variance in SMBH growth.

Consequently, $\tau_{\rm coherence}$ has clear population-level implications. A very short coherence timescale drives SMBHs with similar seed masses and formation times to follow nearly identical, smooth growth tracks, set primarily by their average halo accretion rate. In contrast, a longer $\tau_{\rm coherence}$ induces genuine diversity in growth paths even at fixed halo accretion rate, generating

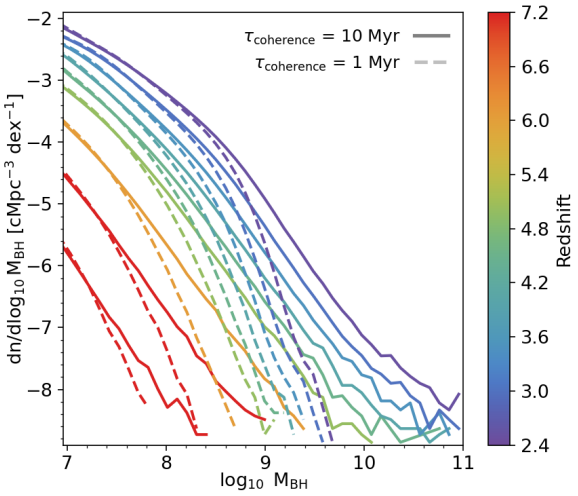

Figure 5.12: Black hole mass function (BHMF) for all SMBHs in the simulation. Solid lines show the fiducial run (Tab. 5.1; $\tau_{\rm coherence} = 10$ Myr) at different redshifts (colors), while dashed lines show the results with a shorter coherence timescale of $\tau_{\rm coherence} = 1$ Myr, keeping all other parameters fixed. The choice of $\tau_{\rm coherence}$ has a strong impact on the high-mass tail of the BHMF, which corresponds to the most massive SMBHs powering bright quasars across cosmic history.

intrinsic mass scatter in addition to the stochasticity already encoded in the distribution $P(\eta_{\rm acc}|s\dot{M}_{\rm cold,acc})$. The latter primarily governs the short-term variability seen in individual quasar light curves.

These effects are illustrated in Fig. 5.12, which shows the black hole mass function (BHMF) for the entire SMBH population at different redshifts. We compare two cases: our fiducial model with $\tau_{\rm coherence} = 10$ Myr and an alternative run with $\tau_{\rm coherence} = 1$ Myr, keeping all other parameters fixed (Tab. 5.1). The contrast is striking: the 1 Myr run produces a much narrower BHMF, making it substantially more difficult to grow the most massive SMBHs observed at all redshifts. As a consequence, the individual growth histories shown in Figs. 5.5 and 5.6 would appear far more uniform for $\tau_{\rm coherence} = 1$ Myr, with significantly reduced diversity in accretion trajectories.

The $M_{\rm BH}$–$M_{\rm halo}$ relation discussed in Sec. 5.3.3.1 is likewise strongly influenced by $\tau_{\rm coherence}$. A shorter coherence timescale greatly reduces the scatter in this relation – forcing SMBH growth tracks to closely follow those of their host halos – and suppresses the stochastic outliers that, in our model, give rise to the brightest quasars across cosmic time. Following Eq. 5.12, and consistent with the central limit theorem, the distribution of SMBH masses at fixed halo mass approaches a narrow log-normal as $\tau_{\rm coherence}$ decreases.

Conversely, a longer coherence timescale preserves extended high-mass tails in the distribution, enabling a subset of SMBHs to reach extreme masses and power the billion-solar-mass quasars observed in the early Universe.

Despite its importance, the coherence timescale of accretion is inevitably degenerate with other parameters that regulate SMBH growth. For instance, a shorter $\tau_{\rm coherence}$ could, in principle, be offset by increasing the scatter in the accretion rate distribution $P(\eta_{\rm acc}|s\dot{M}_{\rm cold,acc})$ – though this freedom is limited, since the distribution is already anchored to the observed shape of the QLF – or by introducing additional variance through seeding or merger prescriptions. What makes the accretion timescale especially compelling, however, is that it can also be constrained through completely independent methods that probe quasar lifetimes and duty cycles (e.g., Martini 2004). For example, proximity-zone measurements in quasar spectra suggest that quasars must typically have been actively accreting for 10^4–10^7 years to produce the observed ionization structures around them (e.g., Eilers et al. 2017), setting a firm lower limit on the accretion timescale. Meanwhile, clustering-based duty cycle estimates provide complementary constraints by measuring how long quasars, on average, remain above a given luminosity threshold (Martini & Weinberg 2001; Haiman & Hui 2001). Taken together, these independent probes elevate $\tau_{\rm coherence}$ from a tunable modeling parameter to a physically interpretable quantity with broad observational implications.

Indeed, from a physical standpoint, $\tau_{\rm coherence}$ can be interpreted as the characteristic timescale of the processes that regulate quasar activity. These processes remain poorly constrained: it is still unclear whether most variability arises from rapid, small-scale fluctuations in accretion flows, or from longer-term, secular changes associated with galaxy and halo evolution, with short-term variability contributing only secondarily (e.g., Alexander et al. 2025). A more general framework than that developed here could, in principle, capture the full hierarchy of variability timescales by parametrizing the stochastic accretion process in terms of, e.g., its power spectral density, thereby quantifying the relative importance of different physical mechanisms. While developing such a framework lies beyond the scope of this work, it represents a promising avenue for future research. Ultimately, by combining the full suite of constraints – proximity zones and clustering-based estimates of quasar lifetimes and duty cycles, instantaneous accretion traced by the QLF, and long-term SMBH growth inferred from the cERDF and local SMBH mass measurements – it may become possible to phenomenologically uncover the processes that govern SMBH evolution across cosmic time.

5.3.3.3 The relative role of mergers and accretion

In Fig. 5.13, we examine the relative importance of mergers and gas accretion in driving SMBH growth. The solid lines show the BHMF from our fiducial run, where both accretion and mergers are included. The dashed lines

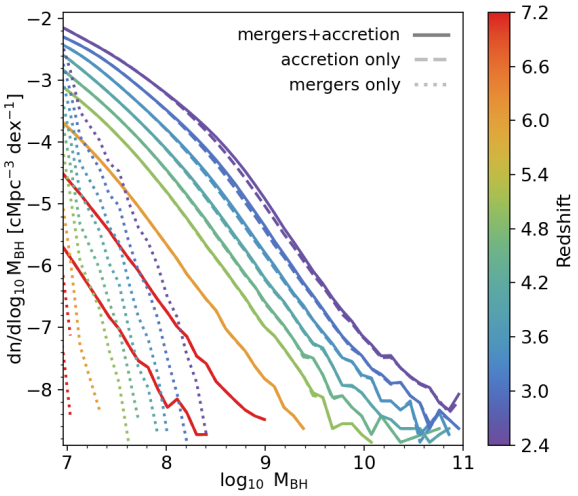

Figure 5.13: Black hole mass function (BHMF) for all SMBHs in the simulation at different redshifts (colors), illustrating the relative contributions of mergers and accretion to SMBH growth in our fiducial model. Solid lines show the full BHMF including both accretion and mergers. Dashed lines show the BHMF when only accretion is included (merging SMBHs are removed without contributing mass), demonstrating that mergers have only a marginal effect on the overall distribution. When accretion is switched off, mergers alone fail to produce sufficiently massive SMBHs to match observations (dotted).

represent a run where mergers are switched off by simply discarding merged SMBHs (i.e., black holes that are "sinked" or disrupted, see Sec. 5.2.1.2) without adding their mass to the remnant. The comparison reveals that mergers contribute only minimally across the entire redshift and mass range probed. At high redshift, the solid and dashed curves are indistinguishable, while at $z \lesssim 4$ a slight difference emerges around $M_{\rm BH} \approx 10^{8.5}$–$10^{9.5}\,{\rm M}_\odot$. These SMBHs likely reside in massive halos where cold-gas accretion has been suppressed; in such cases, mergers provide the only significant growth channel, producing the small offset. However, this difference is marginal and unlikely to affect any of the quasar observables considered here. This conclusion is broadly consistent with previous studies, which suggest that mergers become important only for the most massive SMBHs whose gas accretion has already been quenched (e.g., Shankar et al. 2009; Volonteri 2012; Pacucci & Loeb 2020). Extending our model to lower redshifts will allow us to probe this regime in more detail.

The dotted lines in Fig. 5.13 illustrate the BHMF when accretion is completely switched off and SMBHs grow only through mergers. In this scenario, SMBH masses never exceed $\sim 10^8\,{\rm M}_\odot$, even by cosmic noon. The BHMF evolves from being entirely seed-dominated at high redshift to gradually

incorporating more growing SMBHs, reflecting the slow accumulation of mass as black holes merge hierarchically in the ΛCDM paradigm.

Although mergers do not play a dominant role in shaping SMBH growth or quasar observables in our current framework, quantifying their contribution is nonetheless crucial for an independent and complementary test of SMBH assembly. Gravitational-wave observations, in particular, are sensitive almost exclusively to mergers and provide a window into a regime that is otherwise invisible to traditional electromagnetic probes. Pulsar timing arrays (PTAs) have already begun to place constraints on the gravitational-wave background generated by SMBH binaries (e.g., Agazie et al. 2023), offering indirect evidence for the demographics of massive black hole pairs at $z \lesssim 2$. The upcoming LISA mission, on the other hand, will directly detect individual SMBH merger events over a wide range of redshifts and masses, reaching into the early Universe and probing the very systems responsible for seeding and assembling today's SMBH population (e.g., Amaro-Seoane et al. 2023).

Because our framework is explicitly constructed from cosmological merger trees, it is particularly well suited to generate detailed predictions for the merger rates, mass ratios, and redshift distribution of SMBH binaries. Even if mergers are subdominant for quasar fueling, their gravitational-wave signatures could provide the cleanest observational handle on SMBH assembly histories. In this sense, gravitational-wave observatories will not only test the merger-driven growth channel but also offer an entirely orthogonal way to validate models like ours. In future work, we plan to extend our model in this direction, leveraging its merger-based nature to make concrete predictions for the SMBH merger landscape in the upcoming era of PTAs and LISA.

5.4 Discussion and summary

In this work we introduced BAQARO (Black hole Accretion and Quasar Activity in a Realistic Observational framework), a new empirical model for the cosmological evolution of supermassive black holes (SMBHs) and quasars from cosmic dawn to cosmic noon. The framework is built on subhalo merger trees from the N-body version of the FLAMINGO large-volume simulation (Schaye et al. 2023; Kugel et al. 2023), and links SMBH growth to halo assembly through a compact set of parametric prescriptions designed to capture both average evolutionary trends and stochastic variability. A key design choice is the absence of explicit redshift dependence: cosmic evolution enters naturally through the changing specific halo accretion rate (sHAR), allowing the same physical mapping to be applied seamlessly from the epoch of reionization to cosmic noon. The model produces full SMBH mass growth histories, quasar light curves, SMBH merger trees, and host-halo statistics, providing a versatile platform for direct comparison with a wide range of observational constraints.

The model incorporates three main ingredients – seeding, accretion, and mergers – implemented as follows. (i) Seeding. Because subhalos in our merger trees are only resolved once they reach relatively large masses, we initialize each newly formed halo with a fixed "seed" black hole of mass $M_{\rm start}$ (Sec. 5.2.2.1). This empirical initialization acts as a proxy for unresolved early growth and establishes the baseline of the $M_{\rm BH}$-$M_{\rm halo}$ relation. In this work we adopt a single fiducial value of $M_{\rm start}$, deferring exploration of a distribution of seed masses to future extensions of the model.

(ii) Accretion. We tie the specific black hole accretion rate (sBHAR) to the specific cold halo accretion rate, $s\dot{M}_{\rm cold,acc} = f_{\rm cold}(M_{\rm h})\, s\dot{M}_{\rm acc}$, measured between two consecutive snapshots. The cold fraction $f_{\rm cold}(M_{\rm h})$, taken from Correa et al. (2018), accounts for the suppression of cold inflows in massive halos due to virial shock heating, while allowing efficient accretion in low-mass halos at high redshift. Conditional on $s\dot{M}_{\rm cold,acc}$, we draw the Eddington-normalized accretion rate, $\eta_{\rm acc} = \dot{M}_{\rm BH}/\dot{M}_{\rm Edd}$, from a log-normal distribution whose mean and scatter scale as power laws of $s\dot{M}_{\rm cold,acc}$ (Sec. 5.2.2.3). The radiative efficiency $\epsilon(\eta_{\rm acc})$ is prescribed following slim-disk models (Sądowski et al. 2014; Madau et al. 2014), transitioning from a thin-disk plateau at sub-Eddington rates to a saturated luminosity at super-Eddington rates. This ensures that the bolometric luminosity, $L_{\rm bol} = \epsilon\, \eta_{\rm acc}\, \dot{M}_{\rm Edd} c^2$, remains physically consistent across regimes. To model stochastic variability, SMBH masses are advanced by sub-cycling each snapshot into intervals of a coherence timescale, $\tau_{\rm coherence}$: $\eta_{\rm acc}$ is held constant over $\tau_{\rm coherence}$ and redrawn thereafter. This single parameter controls how strongly growth histories "average out" versus retain long-lived bursts.

(iii) Mergers. When subhalos merge, their central SMBHs are assumed to coalesce following a simplified and optimistic prescription: the remnant SMBH has a mass equal to the sum of the progenitors, and no black hole is ejected from the host subhalo as a result of gravitational recoil (Sec. 5.2.2.2). This treatment is similar to that adopted in many large-scale cosmological hydrodynamical simulations (e.g., Habouzit et al. 2021). We also perform control experiments in which we suppress the mass contribution from mergers, or conversely suppress accretion, to isolate their relative roles.

By construction, BAQARO is anchored to three observational diagnostics that probe complementary aspects of quasar physics (Sec. 5.2.3): (a) the bolometric quasar luminosity function (QLF), which traces the global abundance of quasars as a function of luminosity; (b) the conditional Eddington-ratio distribution function (cERDF), $P(\lambda_{\rm Edd}|L_{\rm bol})$, which leverages broad-line SMBH mass estimates to probe instantaneous fueling at fixed luminosity; and (c) the large-scale clustering of UV-luminous quasars, which constrains typical host halo masses and duty cycles. In practice, we compare our predictions with the bolometric QLF compilation of Shen et al. (2020), cERDF measurements derived from SDSS and high-redshift samples (Wu & Shen

2022; Fan et al. 2023), and the quasar auto-correlation functions from BOSS and SDSS (Eftekharzadeh et al. 2015; Shen et al. 2007) as well as the recent JWST constraints on the high-z quasar-galaxy cross-correlation function (Eilers et al. 2024).

In the analysis presented here, we have focused on the results of a single fiducial run, with the free parameters of the model fixed to the values listed in Tab. 5.1. This calibration was chosen to approximately reproduce the main quasar observables while enabling us to explore the qualitative implications of the framework. In forthcoming work, we will move beyond this fiducial calibration and perform a full Bayesian inference of the model parameters. This will be made possible by developing an emulator trained on the model outputs that can approximate the predicted observables at negligible computational cost (Sec. 5.2.4). The emulator will enable Markov Chain Monte Carlo (MCMC) exploration of the parameter space, allowing us to rigorously quantify parameter degeneracies, assess the constraining power of each observable, and obtain posterior distributions jointly constrained by the QLF, cERDF, and clustering. Such an inference pipeline will sharpen the predictive power of the model, provide robust uncertainty estimates, and establish a systematic connection between phenomenological modeling and observational data.

Our results show that the fiducial model provides a satisfactory match to the bright end of the bolometric QLF ($L_{\rm bol} \gtrsim 10^{45} - 10^{46}\,{\rm erg\,s^{-1}}$), the main evolutionary trends of the cERDF, and the clustering of quasars at $z \approx 2.5$ and $z \approx 6$. Nonetheless, several tensions remain: the model overpredicts the abundance of faint quasars, particularly at high redshift; it yields Eddington ratios that are systematically biased toward slightly higher values than those observed across all redshifts and luminosities; and it underestimates the clustering amplitude at $z \approx 4$, failing to match the strong signal reported by Shen et al. (2007). These discrepancies may point to missing physics in our prescriptions – for example, more complex parametrizations of how accretion is regulated in low-mass SMBHs, or refined treatments of radiative efficiency and luminosity output across different accretion regimes.

At the same time, however, some of these relevant observational constraints remain highly uncertain. The faint end of the QLF is difficult to measure due to incompleteness, obscuration, and contamination from star-forming galaxies. Similarly, the extreme clustering amplitude at $z \approx 4$ is debated, with more recent studies reporting weaker signals (e.g., He et al. 2018). Addressing these issues thus requires advances on both the modeling and observational fronts. Forthcoming wide-field surveys such as DESI, Euclid, and Roman, combined with deep AGN samples from JWST, will provide a far more complete view of quasar demographics and environments, offering critical tests for models like BAQARO.

In addition to confronting the model with key observables, we analyzed the internal assembly of SMBHs in BAQARO and its connection to the growth of their host halos. This leads to several conclusions:

- Accretion dominates SMBH growth. In our model, SMBHs grow predominantly through bursts of near- or super-critical accretion, whereas mergers contribute only a minor fraction of the overall mass budget (Fig. 5.13). Even under optimistic assumptions about merger timescales and remnant survival, the impact of mergers remains marginal. They can provide modest mass boosts for very massive systems ($M_{\rm BH} \sim 10^9\,{\rm M}_\odot$) in gas-poor halos at late times, but they are incapable of producing the billion-solar-mass SMBHs observed as luminous quasars at all redshifts. This result reinforces a broad consensus from both analytical arguments and cosmological simulations that sustained accretion, rather than mergers, is the dominant channel of SMBH assembly across cosmic history (e.g., Shankar et al. 2009; Volonteri et al. 2016).

- Accreting black holes undergo rapid mass assembly at $z \gtrsim 6$, followed by a marked decline in growth rates toward cosmic noon (Fig. 5.5). This overall trend reflects the evolution of halo accretion histories, but stochasticity plays a decisive role in shaping the distribution of SMBH masses. In particular, some black holes become extreme outliers, building up significantly more mass than average through short-lived episodes of very high, but radiatively inefficient, accretion (Fig. 5.6). Such bursts allow rare SMBHs to reach $\sim 10^9\,{\rm M}_\odot$ by $z \approx 6$, consistent with previous models of rapid early SMBH growth (e.g., Madau et al. 2014; Volonteri et al. 2015; Lupi et al. 2016). Crucially, these events are not directly tied to halo mass assembly, but instead emerge from stochastic fluctuations in the accretion rate distribution. Tracking these rare, burst-driven growth histories – rather than focusing solely on population averages – is therefore essential for explaining the emergence of the most massive quasars in the early Universe.

- As a consequence of this SMBH-halo co-evolution, the predicted $M_{\rm BH}$-$M_{\rm halo}$ relation in BAQARO is approximately linear and nearly constant with redshift, with an intrinsic scatter of $\lesssim 0.3$ dex (Fig. 5.11). This tight correlation reflects the average tendency of SMBH growth to follow halo accretion. However, consistent with the stochastic growth episodes discussed above, the most massive SMBHs powering luminous quasars at all cosmic times do not emerge from the mean relation. Instead, they appear as rare outliers, produced by bursts of unusually efficient accretion rather than steady halo assembly. These stochastic extremes are essential for explaining the billion-solar-mass SMBHs observed at

high redshift, but they also complicate simple interpretations of quasar environments that rely on a deterministic $M_{\rm BH}$-$M_{\rm halo}$ mapping.

- A key driver of this stochasticity in SMBH evolution is the coherence timescale of the accretion process, $\tau_{\rm coherence}$. Short values of $\tau_{\rm coherence}$ lead to SMBH mass functions that are narrow, with little diversity in individual growth histories, as accretion fluctuations are averaged out. In contrast, longer coherence timescales preserve broad high-mass tails in the distribution, enabling rare SMBHs to reach $M_{\rm BH} \gtrsim 10^9\,{\rm M}_\odot$ already at early cosmic times (Fig. 5.12). This makes $\tau_{\rm coherence}$ a critical parameter for determining whether the model can produce the most massive quasars seen at $z \gtrsim 6$. Independent constraints from proximity-zone measurements and clustering-based duty cycle estimates suggest timescales for SMBH accretion and quasar activity of $\sim 10^4$–10^7 yr (e.g., Eilers et al. 2017, 2024; Pizzati et al. 2024b), placing $\tau_{\rm coherence}$ in a regime where it directly connects phenomenological modeling with observables. As such, it provides a promising avenue to tie SMBH accretion physics to measurable quantities, and to test whether the observed diversity in quasar activity is consistent with burst-driven growth.

The version of BAQARO presented here is a preliminary implementation of the framework. Several avenues for further development are already clear. First, a full parameter inference must be carried out. This will enable us to quantify parameter degeneracies, identify which observables drive the strongest constraints, and obtain robust posteriors for SMBH growth prescriptions. Such an inference pipeline will significantly strengthen the predictive power of the model.

Second, the redshift range of the model must be extended. Our current analysis is restricted to $2 \lesssim z \lesssim 15$, mainly to reduce computational costs. Extending BAQARO down to $z = 0$ will allow direct tests against the local scaling relations, the observed black hole mass function, and the full history of quasar downsizing. This step will also make it possible to connect high-z accretion-driven growth to the observed demographics of SMBHs in the nearby Universe.

Third, the treatment of the low-mass and seeding regime needs to be improved. At present, the limited resolution of the FLAMINGO simulation prevents us from resolving the formation and early growth of low-mass SMBHs. We plan to tackle this in two complementary ways: (i) by rebuilding the model on the larger, higher-resolution FLAMINGO-10k run (Schaller et al. in prep.), and (ii) by incorporating analytical prescriptions to capture the unresolved early phases of SMBH seeding and growth. Together, these approaches will allow us to explore the critical regime of $M_{\rm BH} \sim 10^4$–$10^7\,{\rm M}_\odot$, which remains one of the most uncertain aspects of SMBH evolution.

With these developments, we hope to address several key questions that remain open in our present analysis. For example: can the strong clustering signal reported by Shen et al. (2007) at $z \approx 4$ be ruled out as an observational systematic, or is there a physically consistent way to connect it with the rapid SMBH buildup at higher redshift and the subsequent evolution to lower redshift? More broadly, how much can we learn from clustering-based estimates of the quasar duty cycle across cosmic time? And how can these constraints be tied to lifetime estimates from quasar proximity zones and damping-wing analyses (e.g., Eilers et al. 2017; Ďurovčíková et al. 2024), and ultimately to the coherence timescale of the accretion process? Because BAQARO resolves individual SMBH accretion histories, it naturally predicts the fraction of time black holes spend above a given luminosity threshold. This definition of duty cycle can be compared directly with clustering-based estimates and with lifetime measurements from proximity zones, providing a coherent test of whether short-lived, bursty accretion episodes are compatible with the observed demographics of quasars.

Another major uncertainty concerns the role of super-critical accretion in the assembly of early SMBHs. Our results suggest that bursts of highly efficient accretion are essential for producing billion-solar-mass quasars by $z \gtrsim 6$, with short-lived episodes of $\sim$1–10 Myr compatible with lifetime and duty cycle constraints at $z \approx 6$ (Pizzati et al. 2024b). Yet it remains unclear how robust this channel is compared to alternative pathways, and to what degree super-critical accretion can complement or replace heavy seeding scenarios. In the current version of the model, SMBHs are initialized with a fixed seed mass, meaning that the degeneracy between seed properties and subsequent accretion histories remains unresolved. Breaking this degeneracy will be critical for distinguishing between different theories of early black hole formation.

Encouragingly, the next generation of observational constraints will provide powerful tests of these ideas. Ongoing and upcoming surveys with JWST, Euclid, and Roman will directly probe the abundance of 10^6–$10^9\,M_\odot$ SMBHs at $z \gtrsim 7$, offering new leverage on the high-redshift early accretion regime. At the same time, a complementary window is opening through gravitational-wave astronomy. Extending the model to the local Universe will allow us to connect with the recent evidence for a nano-Hz gravitational-wave background from pulsar timing arrays – signals that are expected to become even more constraining in the near future (e.g., Agazie et al. 2023). Moreover, the model is well-suited to make forecasts for LISA, which will detect individual SMBH mergers across cosmic time and a broad mass range (e.g., Amaro-Seoane et al. 2023). These gravitational-wave observations will provide an entirely orthogonal test of SMBH assembly, probing the merger-driven channel that is otherwise invisible in traditional electromagnetic surveys.

In summary, our analysis shows that a simple, observationally anchored framework can account for the main demographics of luminous quasars across cosmic time while naturally incorporating the stochasticity required to produce the most extreme SMBHs. Looking ahead, the combination of large-volume simulations, flexible empirical prescriptions, and the rapidly expanding suite of multi-wavelength and multi-messenger observations will enable BAQARO and analogous models to refine our understanding of how the Universe assembled its first quasars and, ultimately, the billion-solar-mass black holes that continue to shape galaxy evolution to the present day.

Acknowledgements

EP is grateful to Victor Forouhar Moreno and Rob McGibbon for help with the HBT-HERONS catalogs. We are grateful to the FLAMINGO team for making their dark matter only simulations available. We acknowledge helpful conversations with the ENIGMA group at UC Santa Barbara and Leiden University. JFH and EP acknowledge support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 885301). This work is partly supported by funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860744 (BiD4BESt). This work used the DiRAC Memory Intensive service (Cosma8) at the University of Durham, which is part of the STFC DiRAC HPC Facility (www.dirac.ac.uk). Access to DiRAC resources was granted through a Director's Discretionary Time allocation in 2023/24, under the auspices of the UKRI-funded DiRAC Federation Project. The equipment was funded by BEIS capital funding via STFC capital grants ST/K00042X/1, ST/P002293/1, ST/R002371/1 and ST/S002502/1, Durham University and STFC operations grant ST/R000832/1. DiRAC is part of the National e-Infrastructure.

6 Towards inference of overlapping gravitational wave signals

Abstract

Merger rates of binary black holes, binary neutron stars, and neutron star-black hole binaries in the local Universe (i.e., redshift $z = 0$), inferred from the Laser Interferometer Gravitational Wave Observatory (LIGO) and Virgo, are 16–130 Gpc^{-3} yr^{-1}, 13–1900 Gpc^{-3} yr^{-1}, and 7.4–320 Gpc^{-3} yr^{-1}, respectively. These rates suggest that there is a significant chance that two or more of these signals will overlap with each other during their lifetime in the sensitivity-band of future gravitational-wave detectors such as the Cosmic Explorer and Einstein Telescope. The detection pipelines provide the coalescence time of each signal with an accuracy $\mathcal{O}(10\,\mathrm{ms})$. We show that using a prior on the coalescence time from a detection pipeline, it is possible to correctly infer the properties of these *overlapping signals* with the current data-analysis infrastructure. We study different configurations of two overlapping signals created by non-spinning binaries, varying their time and phase at coalescence, as well as their signal-to-noise ratios. We conclude that, for the scenarios considered in this work, parameter inference is robust provided that their coalescence times in the detector frame are more than $\sim$ 1–2s. Signals whose coalescence epochs lie within $\sim 0.5\,\mathrm{s}$ of each other suffer from significant biases in parameter inference, and new strategies and algorithms would be required to overcome such biases.

Published in: **EP**, Surabhi Sachdev, Anuradha Gupta, and Bangalore Sathyaprakash. *Toward inference of overlapping gravitational-wave signals*, Physical Review D, vol. 105, no. 10, 2022, doi.org/10.1103/PhysRevD.105.104016
Reprinted here in its entirety.

6.1 Introduction

The advent of the third generation (3G) gravitational-wave (GW) observatories, such as the Cosmic Explorer (CE) (Evans et al. 2021; Reitze et al. 2019a; Reitze et al. 2019b) and the Einstein Telescope (ET) (Punturo et al. 2010), will offer the possibility to observe binary coalescence events from redshifts $z \sim 10$–50, thanks to an order of magnitude improved strain and frequency sensitivity compared to the current generation of detectors of Advanced LIGO (Aasi et al. 2015), Advanced Virgo (Acernese et al. 2015), and KAGRA (Akutsu et al. 2019). Indeed, 3G observatories will have unprecedented sensitivity to detect coalescence events from an epoch when the Universe was still in its infancy assembling its first stars and will routinely detect mergers with stupendously large signal-to-noise ratios of several thousands (Sathyaprakash et al. 2012; Vitale & Evans 2017; Maggiore et al. 2020; Evans et al. 2021). An order of magnitude greater redshift reach and access to extremely high-fidelity signals compared to current interferometers promises many new discoveries, while allowing completely independent, precision tests of cosmological models, alternative gravity theories, and astrophysical scenarios of compact binary formation and evolution (Evans et al. 2021; Maggiore et al. 2020). With an expected rate of hundreds of thousands of binary coalescence signals each year (Baibhav et al. 2019; Sachdev et al. 2020; Maggiore et al. 2020; Evans et al. 2021) on top of weak, but persistent, radiation from isolated neutron stars (Sathyaprakash et al. 2012), rare bursts from supernova and other transient sources and stochastic backgrounds (Regimbau et al. 2017), 3G observatories demand novel algorithms for signal detection and characterization. Therefore, a proper understanding of systematics arising from overlapping loud and quiet signals alike will answer a range of scientific questions that are at the forefront of fundamental physics and astronomy, as well as a realistic estimation of the computational cost.

According to current estimates, 3G observatories are expected to detect hundreds of thousands of binary black hole (BBH) and binary neutron star (BNS) mergers each year (Baibhav et al. 2019; Sachdev et al. 2020; Maggiore et al. 2020; Evans et al. 2021). If we take account of the fact that signals will last longer due to a lower starting frequency (3 Hz for ET and 5 Hz for CE), then it is clear that 3G data will be dominated by many overlapping signals (Regimbau et al. 2012; Meacher et al. 2016; Regimbau et al. 2017; Samajdar et al. 2021; Relton & Raymond 2021). The problem of overlapping signals producing a confusion background in future terrestrial detectors was identified more than a decade ago (Regimbau & Hughes 2009). The problem poses two challenges: first, the detection of individual signals could, in principle, be affected by the presence of multiple signals. Second, the current Bayesian inference methods (Veitch et al. 2015; Ashton et al. 2019) may

not guarantee unbiased estimation of source parameters, which is crucial to deliver the science promises of 3G observatories.

A similar issue has been tackled, in a different context, by the LISA (Laser Interferometer Space Antenna) community. LISA is expected to produce a data set containing many overlapping astrophysical signals: galactic white dwarf binaries are persistent sources of gravitational waves and they produce a "foreground" noise (Crowder & Cornish 2004) that could masquerade the detection and parameter estimation of other astrophysical signals. Several authors have studied the problem of both detection (Cornish & Porter 2007; Littenberg 2011; Babak et al. 2010) and Bayesian inference (Cornish & Crowder 2005; Crowder & Cornish 2007) in this context, while others have focused on searching for the global solution to the full family of potential signals (Littenberg et al. 2020; Robson & Cornish 2017; Petiteau et al. 2013). A parallel effort has been made by other studies (Cornish & Littenberg 2015; Chatziioannou et al. 2021; Cornish et al. 2021) to characterize the overlapping between GW signals and glitches in the context of LIGO/Virgo data analysis. These studies represent a useful reference that could guide the development of new algorithms specifically suited to deal with the parameter estimation of multiple signals in the context of terrestrial detectors.

However, no effort to study the problem of inference in the case of 3G terrestrial detectors has so far been made. Given the relevance of this specific problem, an exploratory study of the capabilities of current parameter estimation methods in the context of overlapping signals in terrestrial detectors appears to be necessary. With this consideration in mind, we aim to characterize the conditions for which parameter estimation is possible with the current algorithms for overlapping signals and to identify regions in the signal parameter space that create significant biases in the inference process, for which novel algorithms would be required.

Detecting overlapping GW signals has been shown to be possible by two ET mock data challenges (Regimbau et al. 2012; Meacher et al. 2016). These studies were able to correctly identify and recover signals even when they were overlapping with multiple others. Even though the signal detection may provide unbiased results, however, there is no guarantee that the parameter inference in the case of overlapping signals is possible within the current framework. This is because current methods heavily rely on the efficiency of sampling algorithms, which are used to explore the posterior distribution of parameters. If we analyze overlapping signals with the current parameter estimation (PE) procedures (i.e., the assumption that the parameter space for multiple signals is the same as in the case of data containing only one signal at a time), we expect Markov Chains and the posterior distribution to exhibit a non-trivial behavior such as slowly or non-convergence of chains, multi-modal and biased posterior distributions, etc.

To this end, we deploy the Fisher information matrix formalism to gauge the limit between the region where overlapping signals could lead to biases in

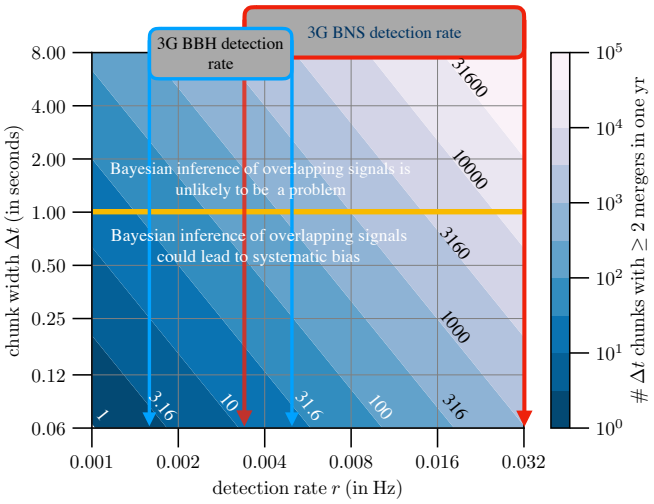

Figure 6.1: Contour diagram showing the number of times two or more signals have their epoch of coalescence occurring within an interval Δt in a year's worth of data as a function of the chunk size Δt and the Poisson rate r. Also shown are the detection rate of BBH and BNS signals in 3G observatories of one ET and 2 CEs (Samajdar et al. 2021). As an example, if the detection rate is 8 mHz then we can expect in one year's of data 1000 one-second long chunks in which two or more mergers would occur. For a pair of signals whose coalescence times differ by an interval of $\Delta t > 1\,\mathrm{s}$ we do not expect to see any biases in their parameter estimation, although the signals themselves might overlap. Biases begin to show up for $\Delta t < 1\,\mathrm{s}$ and become severe as $\Delta t \to 0$.

parameter inference and the region where they don't. The Fisher study tells us that as long as the difference in the merger time Δt_C of two overlapping signals is larger than the accuracy δt with which their merger times can be measured (i.e., $\Delta t_C \gg \delta t$), irrespective of how long the individual signals are, parameter inference will not cause significant biases. We exploit this result in the Bayesian analysis of mock data by choosing the prior on the merger epoch as determined by the signal detection pipelines, which is about $\delta t_C \sim \mathcal{O}(10\,\mathrm{ms})$ (Liting et al. 2014). Indeed, most signals are recovered by search pipelines with a bias of $\delta t_C < 20$ ms. A conservative prior on the merger time could be a factor of 10 to 20 larger or at most 500 ms. Thus, two overlapping signals with their merger times separated by larger than about $\approx 2\,\mathrm{s}$ are not expected to suffer from any systematic biases. Hence, it suffices to consider the extent to which overlapping signals pose a problem for $\Delta t_C \lesssim 2\,\mathrm{s}$.

The rest of the paper is organized as follows: in Sec. 6.2, we compute the number of chunks in a year's worth of data containing more than one merger. Section 6.3 is devoted to studying the covariance between overlapping signals using the Fisher information matrix with the emphasis on what we might

expect for parameter inference in case of overlaps. Bayesian inference of overlapping signals is presented in Sec. 6.4. Our main conclusions and a brief discussion of the type of problems that should be addressed in future studies is presented in Sec. 6.5.

6.2 Number of overlapping signals

The number of overlapping signals depends on (a) the typical duration of signals and (b) the rate at which they arrive at the detector. At the leading order, the length ξ of a coalescing compact binary signal starting from a gravitational-wave frequency f_s until merger is given by

$$\xi = \frac{5}{256} \left(G\mathcal{M}/c^3\right)^{-5/3} (\pi f_s)^{-8/3}, \tag{6.1}$$

where G is Newton's constant, c is the speed of light and the chirp mass $\mathcal{M}$ is related to the component masses m_1 and m_2 via $\mathcal{M} \equiv (m_1\, m_2)^{3/5}/(m_1 + m_2)^{1/5}$. A BNS system consisting of a pair of $1.4\, M_\odot$ would last for $\xi \simeq 10^3$ s starting from a frequency of $f_s = 10$ Hz (relevant for Advanced LIGO and Advanced Virgo), 1.8 hr for $f_s = 5$ Hz (CE) and almost 7 hr for $f_s = 3$ Hz (ET). A source of intrinsic chirp mass $\mathcal{M}$ at a cosmological redshift of z would appear in the detector to have a chirp mass of $(1+z)\mathcal{M}$, and hence lives for a shorter duration in a detector's sensitivity band. Thus, BNSs ($1M_\odot \leq m_1, m_2 \leq 3M_\odot$) could last for tens of minutes to several hours in band while BBH signals ($3M_\odot \leq m_1, m_2 \leq 50M_\odot$) could last for tens of seconds to thousands of seconds.

The cosmic merger rate of compact binary coalescences determined by the first two observing runs of LIGO and Virgo (Abbott et al. 2019b, 2021) implies that in a network of 3G observatories the detection rate r, defined as the number of signals whose matched filter signal-to-noise ratio is larger than 12, lies in the range $r_{\rm BBH} \in [5 \times 10^4, 1.5 \times 10^5]$ yr^{-1} for BBHs and $r_{\rm BNS} \in [10^5, 10^6]$ yr^{-1} for BNSs (Samajdar et al. 2021; Abbott et al. 2018b, 2016b). Thus, given that signals last for several hours, 3G data would contain several loud overlapping signals at any one time. We shall see below that for the purpose of parameter inference the relevant quantity is not how many overlapping signals there are at any one time but if two or more signals have their merger times lie within a duration Δt. This is what we will set out to compute next.

6.2.1 Overlapping signals of the same family

Let r denote the Poisson detection rate of a given signal family (BBH or BNS). In an interval Δt, the expected Poisson rate is $\nu = r\,\Delta t$ and the probability of observing exactly k mergers during Δt is given by

$$P_k(\nu) = \frac{\nu^k\, e^{-\nu}}{k!}. \tag{6.2}$$

Thus, the probability of observing two or more mergers during Δt is

$$P_{k\geq 2} = \sum_{k=2}^{\infty} P_k(\nu) = \sum_{k=2}^{\infty} \frac{\nu^k\, e^{-\nu}}{k!} = 1 - e^{-\nu}(1+\nu). \tag{6.3}$$

We have made use of the fact that the Poisson distribution is normalized, namely $\sum_{k=0}^{\infty} P_k(\nu) = 1$. To compute the number of chunks $N_{k\geq 2}$ in which two or more mergers will be observed we must multiply the probability $P_{k\geq 2}$ by the number of chunks $n_{\Delta t} = T/\Delta t$ in an observational period T:

$$N_{k\geq 2} \equiv P_{k\geq 2} n_{\Delta t} = \left[1 - e^{-\nu}(1+\nu)\right] \frac{T}{\Delta t}. \tag{6.4}$$

Substituting $\Delta t = \nu/r$ and noting that $N_T \equiv r\,T$ is the total number of signals detected during the period T, we get

$$N_{k\geq 2} = \left[1 - e^{-\nu}(1+\nu)\right] \frac{N_T}{\nu}. \tag{6.5}$$

It is easy to see that in the limit $\Delta t \to 0$ (equivalently, $\nu \to 0$), $N_{k\geq 2} \simeq \nu N_T/2$. The factor of $1/2$ assures that the number of instances when two or more signals are found in a chunk is never greater than half of the total number of observed signals but it is also weighed down by the Poisson rate ν. In the other limit, when $\Delta t \to T$ (and $\nu \gg 1$), $N_{k\geq 2} \simeq 1$ but less than 1.

Figure 6.1 plots the number of chunks $N_{k\geq 2}$ in which we can expect to find two or more mergers in a year's worth of data (i.e., using $T = 1\,\mathrm{yr}$ and $\nu = r\,\Delta t$). Also indicated in the plot are the detection rate of BBH (BNS) which is expected to be in the range $r_{\mathrm{BBH}} \in [1.6, 4.8] \times 10^{-3}\,\mathrm{s}^{-1}$ ($r_{\mathrm{BNS}} \in [3.5, 35] \times 10^{-3}\,\mathrm{s}^{-1}$, respectively) (Samajdar et al. 2021) in a 3G detector network comprising of one ET and two CEs (one in north America and the other in Australia). As we shall see in Sec. 6.3, parameter inference should not be a problem if the difference in coalescence times of a pair of signals is larger than $\sim 1\,\mathrm{s}$; this is indicated in Fig. 6.1 by the horizontal line drawn at $\Delta t = 1\,\mathrm{s}$. Thus, in Sec. 6.4 we will focus on Bayesian inference of signals whose merger times differ by about one second. We see that at the higher end of the BNS rate, we expect $\sim 15,000$ one-second long chunks with two or more mergers while at the lower end of the BNS rate this number is ~ 200. Likewise, ~ 300 chunks will contain two or more BBH mergers at

the higher end of the BBH detection rate while this number is ~ 40 at the lower end of the BBH rate. Although the vast majority of events will have their merger times larger than 1 s from their nearest neighbor, the number of events with their merger times within a second is quite large.

The detection rate of BBH signals in the current detector network of LIGO, Virgo and KAGRA at their design sensitivity is at best $r \sim 2.3 \times 10^{-5}$ s^{-1} (or 730 yr^{-1}) (Abbott et al. 2021). Thus, the probability of observing multiple mergers in a chuck of size 1 s or less is negligibly small in the Advanced detector era. This will also be the case in the A+ era (Abbott et al. 2018a) where the detection rates are expected to be 3 times larger.

6.2.2 Overlapping signals from two different families

If the detection rate of signal families A and B are r_A and r_B, then probability that *one* or more mergers of each of these signal families would occur during an interval Δt is

$$P_{A,k\geq 1} = 1 - e^{-\Delta t\, r_A}, \quad P_{B,k\geq 1} = 1 - e^{-\Delta t\, r_B}. \tag{6.6}$$

Thus, the probability P_{AB} that an interval Δt contains one or more from each of the two signal families is simply the product $P_{AB} = P_{A,k\geq 1}\, P_{B,k\geq 1}$. If the rates are small, this reduces to $P_{AB} = (\Delta t)^2\, r_A\, r_B$ and the number of such chunks over a period T is $N_{AB} = (\Delta t)^2 r_A\, r_B\, T = N_A\, N_B / n_{\Delta t}$, where N_A and N_B are the total number of mergers during the period T of families A and B, respectively, and $n_{\Delta t} = T/\Delta t$ is the number of chunks of width Δt during T. Using the range of BNS and BBH rates quoted before, we find that N_{AB} would lie in the range 170–5100 for $T = 1\,\mathrm{yr}$ and $\Delta t = 1\,\mathrm{s}$.

From the foregoing discussions it is clear that a small but significant fraction of signals would have their coalescence time within an interval of 1 s. As we shall see in the next Section, due to their long duration, overlapping BNS signals are far less correlated with each other than overlapping BBH signals. For the same reason, a pair of overlapping BNS and BBH signals are poorly correlated. Hence, in the Bayesian inference problem (Sec. 6.4) we will only consider overlapping BBH signals.

6.3 Covariance among overlapping signals

If two signals are well separated then the covariance between their parameters is zero and we do not expect one signal to affect the parameter inference of the other. As we bring the two signals closer together in time, at some point the presence of one of the signals will begin to bias the estimation of parameters of the other. In this Section we estimate the covariance between the parameters of a pair of overlapping signals using the Fisher matrix

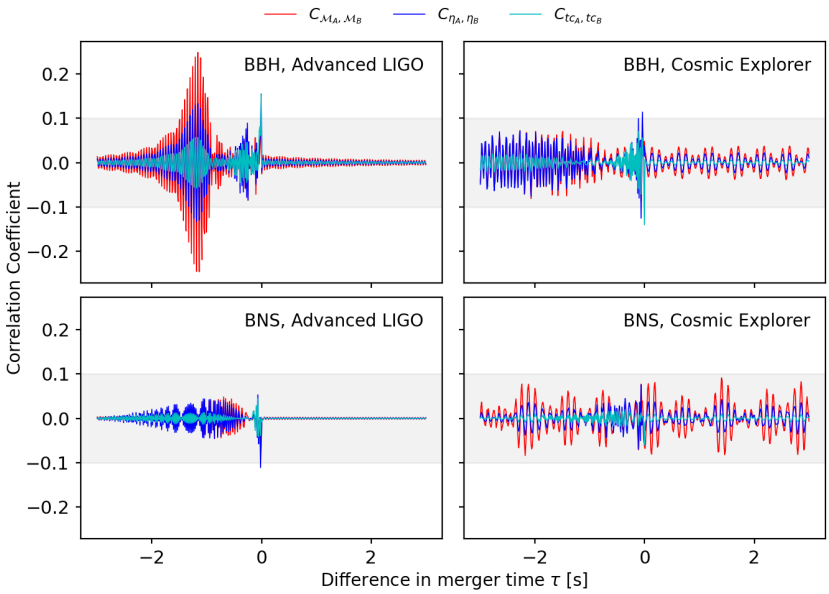

Figure 6.2: Plot shows the correlation coefficients, i.e., normalized covariances as defined by Eq. (6.16) between the parameters of the two overlapping signals as a function of the difference in merger times $\tau = t_C^B - t_C^A$. The left panel is for Advanced LIGO and right for Cosmic Explorer. Top row is for BBHs and bottom row BNSs. We assume the parameter inference of overlapping signals to be negligibly affected when (the absolute value of) the correlation coefficients are less than 10% (grey shaded regions).

formalism. Although Fisher matrix is valid in the limit of large signal-to-noise ratios, any inferences we can draw from the correlation will guide us in choosing the parameter space of compact binaries where systematic biases could be large.

To this end, we assume that the data contains a pair of signals s_A and s_B buried in stationary, Gaussian noise n. The detector output is a sum of the overlapping signals buried in detector noise:

$$x(t) = n(t) + s_A(t, \lambda_\alpha^{(A)}) + s_B(t, \lambda_\alpha^{(B)}). \tag{6.7}$$

where $\lambda_\alpha^{(A)}$, $\lambda_\alpha^{(B)}$, for $\alpha = 1, \ldots, p$, are the set of parameters corresponding to signals s_A and s_B, respectively. Note that since both s_A and s_B are assumed to belong to the same signal family they are specified by the same number of parameters. Furthermore, we shall only consider a single detector for this exercise. The relevant parameters for a binary with non-spinning companions are the chirp mass $\mathcal{M}$, symmetric mass ratio η, the epoch t_C when the signal amplitude reaches its peak and the phase ϕ_C of the signal at that epoch and so: $\lambda_\alpha^{(A)} = (\mathcal{M}^{(A)}, \eta^{(A)}, t_C^{(A)}, \phi_C^{(A)})$ and similarly for signal s_B. We assume the IMRPhenomPv2 waveform model.

For the computation of the covariance matrix it is more convenient to consider that the data contains only one signal, i.e., the sum of the two signals $s = s_A + s_B$, and it is characterized by a double number of parameters: $\theta_a = \lambda_a^{(A)}$ for $a = 1, \ldots, p$ and $\theta_a = \lambda_{a-p}^{(B)}$ for $a = p+1, \ldots, 2p$. For a noise background that is stationary and Gaussian the covariance matrix C, which is inverse of the Fisher matrix Γ, is given by:

$$C_{ab} = \Gamma_{ab}^{-1}, \quad \Gamma_{ab} = \left\langle \frac{\partial s}{\partial \theta_a}, \frac{\partial s}{\partial \theta_b} \right\rangle. \tag{6.8}$$

Here the scalar product of two waveforms (or any pair of functions of time for that matter) h and g is defined as

$$\langle h, g \rangle \equiv 4\Re \int_{f_{\rm low}}^{f_{\rm high}} \frac{\tilde{h}(f)\, \tilde{g}^*(f)}{S_h(f)}\, \mathrm{d}f, \tag{6.9}$$

where $\Re$ stands for the real part of the integral, $\tilde{h}$ and $\tilde{g}$ are the Fourier transforms of the signals h and g, respectively, g^* denotes the complex conjugate of g and $S_h(f)$ is the one-sided noise spectral density of the detector. In our study we will use either the noise spectral density of Advanced LIGO (Aasi et al. 2015) or that of the Cosmic Explorer (Reitze et al. 2019b). The lower frequency cutoff $f_{\rm low}$ is chosen to be 20 Hz for Advanced LIGO and 5 Hz for Cosmic Explorer. For BNSs, the upper frequency cutoff $f_{\rm high}$ is assumed to be the larger of the inner-most stable circular orbit frequency of the two overlapping signals, i.e., $f_{\rm high} = \max[(6^{3/2}\pi M_1)^{-1}, (6^{3/2}\pi M_2)^{-1}]$, where M_1 and M_2 are the total mass of the two overlapping signals. For

BBHs, the upper frequency cutoff is chosen to be the Nyquist frequency of 1024 Hz.

The Fisher matrix contains interference terms of the following type:

$$\Gamma_{\alpha,\,\beta+p} = \left\langle \frac{\partial s_A}{\partial \lambda_\alpha^{(A)}}, \frac{\partial s_B}{\partial \lambda_\beta^{(B)}} \right\rangle . \tag{6.10}$$

Covariances are of primary interest in this Section as they can tell us the degree to which the presence of one signal affects the parameter inference of the other. In order to measure the extent of covariance we consider two sets of overlapping signals (masses are specified in the detector frame):

1. overlapping BBHs with masses:

$$(m_1^{(A)},\, m_2^{(A)}) = (21\,M_\odot,\, 15\,M_\odot) \tag{6.11}$$
$$(m_1^{(B)},\, m_2^{(B)}) = (33\,M_\odot,\, 29\,M_\odot). \tag{6.12}$$

2. overlapping BNSs with companion masses:

$$(m_1^{(A)},\, m_2^{(A)}) = (1.45\,M_\odot, 1.35\,M_\odot) \tag{6.13}$$
$$(m_1^{(B)},\, m_2^{(B)}) = (1.50\,M_\odot, 1.40\,M_\odot). \tag{6.14}$$

Furthermore, in all cases we choose

$$(t_C^{(A)}, \phi_C^{(A)}) = (0,\, 0), \quad (t_C^{(B)}, \phi_C^{(B)}) = (\tau,\, 0), \tag{6.15}$$

and vary τ over the range $[-3,\, 3]$ s.

The covariances between the chirpmass, symmetric mass ratio and epoch of coalescence are plotted in Fig. 6.2 as a function of the parameter τ for overlapping BBHs (top panels) and BNSs (bottom panels) for noise spectral densities of Advanced LIGO (left panels) and Cosmic Explorer (right panels). Other cross-covariances are negligibly small and not shown. What we plot are the normalized covariances, i.e., a combination of the correlation coefficients defined as:

$$\sigma_{ab} \equiv \frac{C_{ab}}{\sqrt{C_{aa} C_{bb}}}, \quad a \neq b. \tag{6.16}$$

This quantity is strictly bounded between -1 and $+1$. A correlation coefficient of $+1$ implies that the parameters are perfectly correlated, -1 implies they are perfectly anti-correlated, and a value of 0 would imply they are uncorrelated. We will take $\sigma_{ab} \sim 0.1$ (grey shaded region in the plot) to be small enough to indicate that the presence of the second signal does not significantly bias parameter inference of the other signal. This threshold is inevitably arbitrary, as a thorough analysis of the connection between the values of

the correlation coefficients and the presence of biases in parameter inference is beyond the scope of this paper. However, as we show in Sec. 6.4.4, the regions of the parameter space where biases in PE arise are compatible with the ones for which $\sigma_{ab} \gtrsim 0.1$.

The correlations displayed in Fig. 6.2 show a range of different behaviours. In all cases, they have a peak for $|\tau| \lesssim 0.5\,\mathrm{s}$. This is expected, as the interaction between the signals is enhanced when the two signals coalesce very close to each other. For $|\tau| > 0.5\,\mathrm{s}$, all the different configurations stay always below the threshold $\sigma_{ab} = 0.1$, with the significant exception of BBH in Advanced LIGO detectors. In this latter case, correlations remain very high in the range $-1.5\,\mathrm{s} < \tau < 0\,\mathrm{s}$, and become small only for $\tau \lesssim -2\,\mathrm{s}$. The fact that correlations are not symmetric in τ can be easily explained by the different form of the two signals considered (see also Fig. 6.3).

Finally, we note that in the case of BNS, the correlation remains always below the threshold both in Advanced LIGO and Cosmic Explorer, except when $\tau \simeq 0$. This implies that parameter inference of overlapping BNS signals is likely to be less severe than that of overlapping BBH signals. We will, therefore, consider only the latter class of signals in the remainder of this paper, leaving the parameter estimation of overlapping BNS signals for future work.

The analysis presented in this section is limited by the fact that we have explored only for a few particular sets of source parameters. Therefore, we cannot conclude that parameter estimation will never be a problem in the case of overlapping BNSs. Indeed, very similar values of the chirp masses (as well as other relevant parameters) will likely increase the correlation between the two signals, especially in the proximity of $\tau = 0\,\mathrm{s}$.

In addition, we note that further work is necessary to assess the validity of the correlation threshold we have considered here, especially in light of the fact that sinusoidal features with amplitudes $\sigma_{ab} \approx 0.05 - 0.1$ are present in the case of the Cosmic Explorer detector, even for large values of $|\tau|$. Despite the fact that these correlations are very low, their effects on the results of parameter inference need to be evaluated quantitatively.

6.4 Bayesian inference of overlapping signals

In this Section, we support the results we have derived using the Fisher information matrix formalism (Sec. 6.3) with a full Bayesian inference procedure. With this *parameter estimation* (PE) process, we are able to fully explore the posterior distribution of the parameters that generated the signals. This is important, because it allows us to confirm the presence (expected from the Fisher study) of distinct maxima in the posterior, one for each signal coalescing within the time chunk considered. Moreover, thanks to this numerical approach, we can explore more carefully the region where

biases are expected assessing their significance and gauging the conditions for which they seem to happen.

Within the Bayesian framework, given a set of parameters λ describing a compact binary coalescence (CBC) waveform $h(\lambda, t)$, we can write the posterior distribution for λ as:

$$P(\lambda\,|x, h) = \frac{\pi(\lambda)\,\mathcal{L}(x\,|\lambda, h)}{\mathcal{Z}(x)}, \tag{6.17}$$

where x is the detector output. This posterior can be explored by using a sampling algorithm (e.g., MCMC, nested sampling). As in Sec. 6.3, assuming that the data x contains two overlapping signals s_A (*signal A*) and s_B (*signal B*), then it can be written as:

$$x = n + s_A + s_B, \tag{6.18}$$

where n is the noise of the interferometer. Note that, in principle, to perform a Bayesian analysis of two or more overlapping signals we should broaden the parameter space, e.g., $\theta = \{\lambda^A, \lambda^B\}$, in order to account for the presence of multiple overlapping signals. However, since running a sampling algorithm requires significant amount of computational resources, in most cases this is not required. In fact, as argued in Sec. 6.3, if the signals' coalescence times are wide apart we do not expect the presence of one signals to influence posterior distribution of parameters of the other. For this reason, in what follows we consider the parameter space of a single CBC signal. We will return to this point later on when discussing possible biases arising because of this choice.

6.4.1 Choice of signal families

As already mentioned, in this analysis we focus only on BBH signals. This choice is motivated by the fact that: (a) covariances among overlapping BNS signals are smaller than the BBH ones (Sec. 6.3), and, therefore, biases in the BNS case are expected to be less important; (b) BNS signals last for several hours in 3G detectors and tens of minutes in Advanced LIGO and Virgo, implying that Bayesian inference takes a formidable amount of computational resources (although new algorithms are already showing the promise of greatly reducing the computational requirement (Zackay et al. 2018; Cornish 2010; Finstad & Brown 2020)).

Furthermore, we also restrict our analysis using Advanced LIGO sensitivity. As argued before, LIGO is not affected by the problem of overlapping signals, because the rate and the duration of the signals are far too small to create any overlap. Nonetheless, in this work we are not really interested in reproducing a realistic set of overlapping data; instead, we want to focus on the parameter estimation process. To do so, there is no substantial

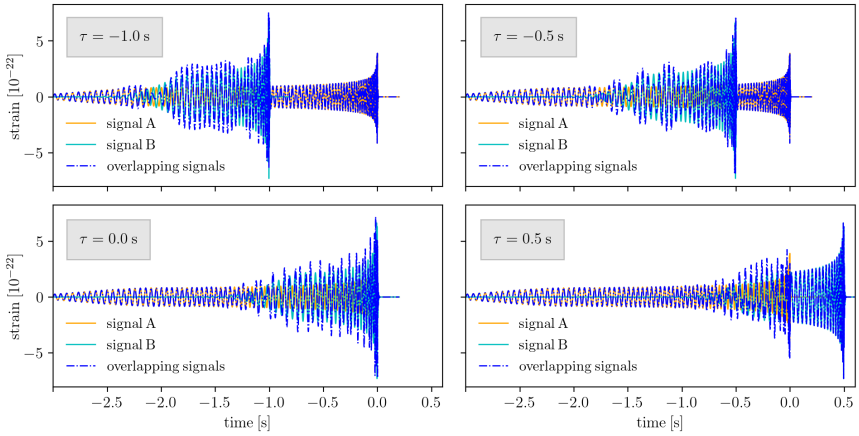

Figure 6.3: Signals in the time domain, for four different values of the time shift τ. *signal A*, *signal B*, and the resulting overlapping signal are plotted. The waveforms are generated using the approximant IMRPhenomPv2. The two luminosity distances are fixed to $d_L^{(A)} = 1\,\mathrm{Gpc}$, $d_L^{(B)} = 1\,\mathrm{Gpc}$. Note that, if we neglect the effects of cosmological redshift, then changing these distances just results on a scaling of the signals' amplitude.

advantage in using 3G mock data: we expect that our conclusions will be valid even if they are based on the analysis of Advanced LIGO mock data.

The parameters of the overlapping BBH signals used in Bayesian inferences is the same as what we used in Sec. 6.3: nonspinning BBHs with masses as given in Eq. (6.12) and coalescence times and phases as given in Eq. (6.15). We ignore the position of the sources in the sky and their orientation relative to the detectors (setting all angles to zero). We do, however, include in our analysis the luminosity distance d_L of the source. The parameter space we use in our analysis is thus:

$$\lambda = \{m_1, m_2, \phi_C, t_C, d_L\}$$

Note that our choice of sky position is the worst case scenario, because we are considering the two sources to have the same exact location in the celestial sphere. In reality, if overlapping signals arrive from different directions in the sky, they will have different phase coherence amongst a network of detectors and thus easier to discriminate. Thus, since our choice of sky position is the worst case scenario, the parameter estimation problem can only be better when sky position and orientation are taken into account.

To explore different configurations of the parameters, we vary the time shift τ - defined in Eq. (6.15) as the epoch coalescence of *signal B* - in the range $\tau \in \{-2.0\,\mathrm{s}, 0.5\,\mathrm{s}\}$. Along with the time shift, we also vary the two luminosity distances of the sources $d_L^{(A)}$ and $d_L^{(B)}$, and their phases $\phi_C^{(A)}$ and $\phi_C^{(B)}$. In the first set of runs, we fix $\phi_C^{(A)} = \phi_C^{(B)} = 0$ and vary the two

distances. We keep the distance of one of the sources fixed to 1 Gpc and set the other at either 500 Mpc, 1 Gpc, or 2 Gpc. In the second set of runs, we vary the phase of signal B ($\phi_C^{(B)} \in \{0, \pi/3, 2\pi/3\}$), keeping $\phi_C^{(A)}$ fixed to zero and the two luminosity distances to $d_L^{(A)} = d_L^{(B)} = 1\,\mathrm{Gpc}$.

The resulting variations in the parameter sets are:

$$\tau = \{-2.0\,\mathrm{s}, -1.5\,\mathrm{s}, -1.0\,\mathrm{s}, -0.5\,\mathrm{s}, 0.0\,\mathrm{s}, 0.5\,\mathrm{s}\} \quad (6.19)$$

$$d_L^{(B)},\, d_L^{(A)} = \{500\,\mathrm{Mpc}, 1\,\mathrm{Gpc}, 2\,\mathrm{Gpc}\} \quad (6.20)$$

$$\phi_C^{(B)} = \{0, \pi/3, 2\pi/3\} \quad (6.21)$$

With these choices, there are 48 different possible configurations, each of which is analyzed for Bayesian parameter inference.

In the inference problem we use a signal model that accurately represents the BBH waveforms. As in Sec. 6.3, we use the IMRPhenomPv2 approximant to create waveforms in the frequency domain, fixing the low frequency cutoff to be 20 Hz, which is consistent with the minimum frequency used in the LIGO/Virgo PE. In Fig. 6.3, we plot the two waveforms in the time domain, for the different configurations of the time shift τ. The resulting overlapping waveform is plotted as well. In Table 6.1, we compute the expected matched filter SNR for the different possible configurations of the parameters, focusing on the distances, since neither the coalescence time nor the phase affect the SNR value.

Table 6.1: SNRs for the two signals we have chosen to focus on in our analysis (considering the two LIGO interferometers network), created with different values of the luminosity distances d_L. Note that applying a time shift to the signals do not change the value of the SNR.

SNR	$d_L = 0.5\,\mathrm{Gpc}$	$d_L = 1\,\mathrm{Gpc}$	$d_L = 2\,\mathrm{Gpc}$
signal A	54.2	27.1	13.5
signal B	82.8	41.5	20.7

6.4.2 Setting up Bayesian inference runs

Having created the mock data with overlapping signals we next focus on parameter inference. Our analysis uses two LIGO interferometers, but our conclusions are not significantly affected by this choice: considering a different detector network would simply result in different SNRs for the signals as we are not focusing on the sky position of the source. Although this could in principle change the heights of the peaks in the posterior distribution, we do expect it to influence their relative ratios significantly, and hence the PE process we consider is expected to hold for any network.

The data set consists of 4 s of mock data from the two LIGO interferometers. 4 s is large enough to span the full length of the longer signal. We do not add any noise to the data – i.e., we set $n = 0$ in Eq. (6.18) –, as we want to highlight the presence of biases created by the overlap between the signals, and these biases could be covered by the statistical uncertainty created by the presence of noise.

We use the BILBY package to perform Bayesian parameter inference of the two signals, running the DYNESTY sampler (Speagle 2020). DYNESTY is a dynamic, nested sampling algorithm (Skilling 2006; Higson et al. 2018), which is well suited for our purposes because it quickly achieves convergence, but at the same time it is able to handle non-trivial, multi-modal distributions better than MCMC-based algorithms (Speagle 2020). We allow the sampler to explore the likelihood surface with respect to all the parameters except ϕ_C, over which the likelihood is analytically marginalized, and d_L, over which the likelihood is numerically maximized. Marginalization over ϕ_C and d_L correctly accounts for the effects of the parameters ϕ_C and d_L on the resulting 3-d posterior (Veitch et al. 2015; Singer & Price 2016).

6.4.3 Bayesian priors

At the beginning of the analysis, we have to set the priors on the various parameters. We consider a uniform prior on the phase ϕ_C, with periodic boundary conditions, a power-law prior on the luminosity distance, $p(d_L) \propto d_L^\alpha$ with $\alpha = 2$, and a uniform prior on the two masses m_1 and m_2 over the range $[10\,\mathrm{M}_\odot, 50\,\mathrm{M}_\odot]$. As for the coalescence time, selecting the best possible prior turns out to be a game-changing strategy. In fact, running a simulation with a wide prior on the time t_C that spans the merger times of the two overlapping signals leads to significant problems: while one of the two signals is always recovered correctly, the other is completely ignored by the sampling algorithm. A wide prior on t_C, therefore, would only allow us to infer the parameters of the louder signal, without access to the weaker one.

However, as already pointed out, previous work suggests that signals can always be detected, even if they are overlapping, and their merger time correctly identified (Regimbau et al. 2012; Meacher et al. 2016). Although these studies dealt only with BNS signals, we do expect that similar conclusions hold also in the case of BBH. This is because (as we show in Sec. 6.4.4, Fig. 6.5) biases on the values of t_c recovered from our PE analysis are minor (at the ms level) and the presence of the overlap does not seem to hamper the time recovery of the signals. However, future efforts will need to back up this assumption and confirm that BBH overlapping signals can be correctly recovered in time domain. From current pipelines, we know that the detection of a signal allows us to know its epoch of coalescence with very low uncertainty (at the order of 10 ms). We then assume to know the time

of coalescence of the two overlapping signals with a good degree of accuracy, and constrain our parameter space choosing a prior on the coalescence time which is centered on the (fiducial) true value of the time t_C, with a width of 100 ms. In this way, for each of the signals we can isolate the region of the parameter space where we expect to find the true values of the injection parameters. This choice allows us to recover the correct parameters for both *signal A* and *signal B*.

Therefore, for each of the 48 injections, we run the Bayesian inference procedure two times: the first one (we refer to it as *run A*) aims to recover the true values of the parameters of *signal A*; to this end, since $t_C^{(\mathrm{A})} = 0.0\,\mathrm{s}$, we set the prior on the coalescence time centered around zero. *Run B*, on the other hand, focuses on the *signal B* peak in the parameter space; thus, the prior is chosen to be centered in $t_C = \tau$.

6.4.4 Results

In this section, we study the posterior distributions obtained for the different runs described in Sec. 6.4.1 and we compare them with the same results obtained when only a single signal is present in the data. This comparison allows us to assess the presence of biases created by the overlap of the signals. In this analysis, we focus on the results for the two masses m_1 and m_2 (which we can rewrite also as chirp mass $\mathcal{M}$ and mass ratio q), and for the coalescence time t_C.

We start by plotting four different corner plots for specific values of the parameters (Fig. 6.4). In the top row, we show the posterior distributions for *run A* (left panel) and *run B* (right) for the following parameters: $d_L^{(A)} = d_L^{(B)} = 1\,\mathrm{Gpc}$; $\phi_C^{(A)} = \phi_C^{(B)} = 0$; $\tau = -1.0\mathrm{s}$. The blue contours represent the results obtained when the two signals are overlapping, while the green ones are the results for a run where only signal A (B) is present in the data. The agreement between these two posteriors (upper panels) is remarkably good, and biases, if any, are negligible. The recovered values of the parameters in the case of *run A* (*run B*) are perfectly compatible with the injected ones $\lambda^{(A)} = \{m_1 = 21\,\mathrm{M}_\odot, m_2 = 15\,\mathrm{M}_\odot, t_C = 0.0\,\mathrm{s}\}$ ($\lambda^{(A)} = \{m_1 = 33\,\mathrm{M}_\odot, m_2 = 29\,\mathrm{M}_\odot, t_C = \tau\}$). This proves that using the current parameter inference methods to deal with overlapping signals is possible.

These results also imply that the posterior distribution for a run with wider priors would be (at least) bi-modal, as the two peaks identified by the two runs (corresponding to the true values of the parameters λ_A and λ_B) with narrower priors would be preserved when the priors are extended coherently to a larger parameter space. However, as already mentioned in Sec. 6.4.3, when we try to extend the prior on the time shift τ, we find that the sampling algorithm can identify only one peak in the posterior. This

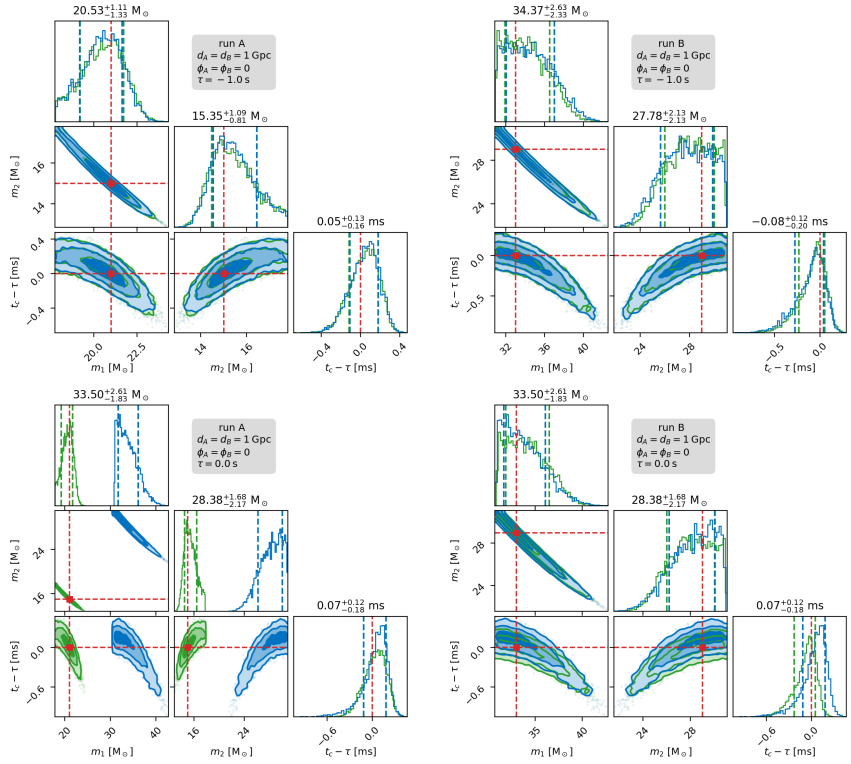

Figure 6.4: Corner plots for two *runs A* (left side) and two *runs B* (right side); all the overlapping signals are created with the following choice of the parameters: $d_L^{(\mathrm{A})} = d_L^{(\mathrm{B})} = 1\,\mathrm{Gpc}$, $\phi_C^{(A)} = \phi_C^{(B)} = 0$. The top row shows the case $\tau = -1.0\,\mathrm{s}$, while the bottom one shows $\tau = 0.0\,\mathrm{s}$. The three parameters considered here are the two masses m_1 and m_2, and the coalescence time t_C. The true values of these parameters are highlighted with red dashed lines in the corner plots. The blue histograms refer to the actual runs, while the green ones are shown for comparison and they are obtained by injecting only one signal in the data. The dashed vertical lines represent the 1σ error on the parameters. On top of each panel, the median values (and the 1σ errors) of the parameters are shown.

behavior is due to the fact that the heights of the two peaks differ by many orders of magnitude, since the peak of $\log \mathcal{L}$ scales as the SNR squared, and the SNRs for signal A and signal B are $\mathrm{SNR}^{(A)} = 27.1$ and $\mathrm{SNR}^{(B)} = 41.5$, respectively (see also Table 6.1). Clearly, the sampling algorithm is not able to sample such a subdominant peak in the posterior. Thus, setting the appropriate prior on the coalescence time t_C, as determined by the search pipeline, is critical in determining the parameters of both of the signals.

We note that a different approach could consist of imposing narrower priors on the two masses m_1 and m_2 (or, equivalently, on the chirp mass $\mathcal{M}$) in order to isolate one peak and exclude the other. This is also a viable alternative, provided that the information on the masses recovered from the detection pipeline is accurate enough to give effective constraints for the priors. Ultimately, combining the information on the coalescence time with the one on the masses may be the best strategy in order to isolate the two peaks even when the two signals are coalescing very close to each other. It is, however, important to ascertain the extent to which such constraints can imposed by carrying out the detection problem on a large sample of injections and the accuracy with which detection pipelines are able to measure chirp mass.

In fact, our approach fails when the two signals are overlapping within 100 ms. In the bottom row of Figure 6.4, we show exactly this case: we take the same distances and phases as described above, but we impose a zero time shift between the two signals. Therefore, in this case the two runs *run A* and *run B* yield the exact same results (as both the priors and the likelihood are the same). As expected, only the louder signal (i.e., signal B) is correctly recovered, with the posterior distribution resembling very closely (although not perfectly matching) the one obtained in the single signal case. We conclude that, once again, the bias is negligible for *run B*. As for signal A, the peak corresponding to $\lambda^{(A)}$ is completely neglected by our inference pipeline, and thus there is no way we can reconstruct the parameters of signal A correctly. This is an intrinsic limitation of our method: different inference prescriptions need to be devised in order to deal with the case of closely coalescing signals.

6.4.4.1 Dependence on the luminosity distance

We now analyze the results of the other runs, where we changed the time shift, luminosity distance, and phase of coalescence of the two signals (as described in Sec. 6.4.1). The top row of Fig. 6.5 shows the posterior distributions for the chirp mass $\mathcal{M}$, the mass ratio q, and the coalescence time t_C, for different combinations of luminosity distances $d_L^{(A)}$, $d_L^{(B)}$ and coalescence times t_C; the phase at coalescence of the two signals are set to $\phi_C^{(A)} = \phi_C^{(B)} = 0$.

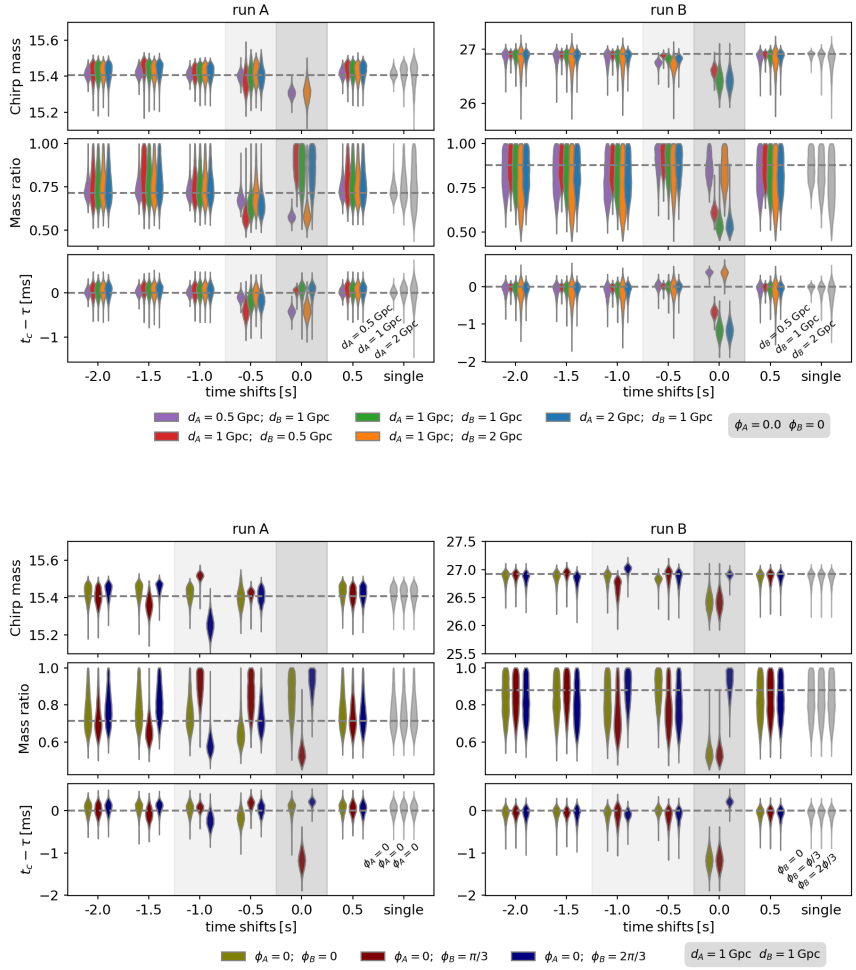

Figure 6.5: Summary of the results for the set of 48 runs, each one with a different configuration of the parameters τ, $d_L^{(\mathrm{A})}$ and $d_L^{(\mathrm{B})}$ (top panels), $\phi_C^{(A)}$ and $\phi_C^{(B)}$ (bottom); for details about the parameters choice, see Sec. 6.4.1. *A runs* are shown on the left panels, and *B runs* are on the right panel. Posterior distributions for the chirp mass $\mathcal{M}$, mass ratio q, and coalescence time $t_C - \tau$ are shown in the form of violin plots. Along with the results for overlapping signals, posteriors for the "*single signal*" case (i.e., only one signal is present in the data) are shown in the rightmost side of each panel in grey. The true values of the masses and times for *signal A* and *signal B* are highlighted with dashed horizontal lines. Note that the distributions in the plots referring to the same time shift τ are slightly shifted with respect to their exact value of τ so that they do not overlap with each other. The $\tau = 0.0\,\mathrm{s}$ runs are highlighted with a dark grey shadowed band; other regions where non-negligible biases are present (see discussion in Sec. 6.4.4) are highlighted in the same way with a lighter shade of grey. Note that in the $\tau = 0.0\,\mathrm{s}$ case, part of the recovered values for the chirp masses are out of the range and thus not shown.

Posteriors are shown in the form of violin plots, and the results for a single injection are shown in light grey color for reference on the right side of each panel. In order to make the plots more accessible, we identify three different regions, highlighted by the shaded grey boxes. In the first region (no shade), biases are negligible: posteriors for *run A* (*run B*) closely resemble the ones obtained by injecting only one signal with the same luminosity distance $d^{(A)}$ ($d^{(B)}$). In this region, the presence of overlapping signal does not create any biases to parameter inference, and both signals can be recovered correctly. As expected from our Fisher analysis (Sec. 6.3), this happens when the two signals are not coalescing too close to each other. In particular, we find that parameter inference is robust in the regions $t_C \lesssim -0.75\,\mathrm{s}$ and $t_C \gtrsim 0.25\,\mathrm{s}$. Note that the asymmetry of these boundary values are expected, as the correlation between the two signals is not symmetric in τ (Fig. 6.2).

When $t_C = -0.5\,\mathrm{s}$ (light shaded region), we find that small biases (at the $1-2\sigma$ level) arise: this implies that the presence of the overlap causes a shift of the posterior peak in the parameter space, preventing the correct recovery of the true parameters $\lambda^{(A)}$ and $\lambda^{(B)}$ for the two signals. We note, however that these relatively small biases may not be a problem in reality, because the presence of the noise may create even larger biases, making these effects totally irrelevant. This depends, of course, on the noise level in the interferometer.

It is also interesting to note that intensity of the biases vary with the relative strengths of the two signals (which are determined by the luminosity distances). In particular, biases for run A (run B) are smaller whenever signal A (signal B) is louder: this can be observed in the left (right) panel of Fig. 6.5, top row, as the posteriors colored in yellow and purple (red, blue, and green) are closer to the ones obtained in the case of a single signal.

Finally, in the last region (darker shade, $\tau = 0.0\,\mathrm{s}$), two relevant effects take place at the same time. First, as already discussed, only the parameters of the louder signal can be recovered correctly. Since the results for *run A* and *run B* are perfectly identical (because they have identical settings), this implies that chirp masses are close to the one of signal A for the yellow and purple cases (as seen in the left panel), and close to the one of signal B in the red, blue, and green cases (as seen in the right one). On top of that, we note that even the louder signal seems to suffer from significant bias in the $\tau = 0.0\,\mathrm{s}$ case. This is again expected from our Fisher analysis (Fig. 6.2), as the correlations between the signals have a peak at zero time shift.

6.4.4.2 Dependence on the phase

In the bottom row of Fig. 6.5, we show the results for the runs with varying ϕ_C. As described in Sec. 6.4.1, we fix the luminosity distances to $d_L^{(\mathrm{A})} = d_L^{(\mathrm{B})} = 1\,\mathrm{Gpc}$ and the phase at coalescence of signal A to $\phi_C^{(\mathrm{A})} = 0$,

and vary $\phi_C^{(\mathrm{B})}$ in the set $\phi_C^{(\mathrm{B})} = \{0, \pi/3, 2\pi/3\}$. Results are presented in the same form as the top row of Fig. 6.5 (Sec. 6.4.4.1).

We find that the phase at coalescence plays an important role in determining whether inference suffers from significant biases or not. In particular, biases are greater for the two configurations $\phi_C^{(\mathrm{B})} = \pi/3$ and $\phi_C^{(\mathrm{B})} = 2\pi/3$. On top of that, they extend in a larger time span: the region where 2σ biases are present extend out to $\tau = -1.0\,\mathrm{s}$; for $\tau = -1.5\,\mathrm{s}$ and $\tau = -2.0\,\mathrm{s}$, they progressively become less severe until they become hardly detectable. Again, we find that biases arise only for negative values of the time shift τ, in accordance with the asymmetric correlation amplitudes found in Fig. 6.2.

Overall, our Bayesian inference analysis confirms the results we found in Sec. 6.3 for BBH in Advanced LIGO detectors (Fig. 6.2, upper left panel). If the two BBH signals do not coalesce too close in the time domain (i.e., their coalescence times are separated by more than $\approx 1.5\,\mathrm{s}$), then inference results are robust: two distinct peaks are present in the posterior, and they can be well-sampled if a suitable prior on the coalescence time is chosen. This is an interesting conclusion, as the vast majority of BBH signals are expected to belong to this category: from Fig. 6.1, we can estimate that only 0.01% of the signals are expected to coalesce within 1 s.

When the BBH signals do coalesce very close to each other ($|\tau| \lesssim 1\,\mathrm{s}$), though, biases at the $2 - 3\,\sigma$ level may arise, as the correlation between the two signals increases. These biases become even more dramatic as the time shift approaches to zero.

6.5 Discussion and Outlook

We presented a Bayesian inference analysis in the case of overlapping gravitational waves signals. Our goal was to assess the capabilities of current Bayesian inference infrastructure to handle the non-trivial case of one or multiple overlaps taking place within a data segment. This problem is destined to play a major role in 3G detector planning, since the dramatic increase in sensitivity will result in a great number of signals coalescing within a few seconds.

We started from a study based on the Fisher matrix formalism, in which we analyzed the correlation between two overlapping signals. In this way, we were able to determine whether in some regions of the parameter space the overlapping signals were strongly correlated with each other, thus preventing a distinct inference procedure for one signal at a time. We found that BNS signals are less strongly correlated, and that their inference will likely be a problem only for coalescence times really close to each other (at the $10 - 100\,\mathrm{ms}$ level). BBHs, instead, suffer from the presence of a correlation starting from a much greater time shift τ (i.e., the difference between the

two coalescence times). In particular, in the Advanced LIGO BBH scenario, correlations are significant up until $|\tau| \approx 2\,\mathrm{s}$.

We investigated these issues further with a full Bayesian analysis of the two overlapping BBHs. The analysis used the DYNESTY sampling algorithm to describe the posterior distribution for the parameters considered. We showed that, in order to sample a single peak without worrying for the presence of the other one, a possible solution is to impose a narrow prior around the fiducial value (provided by the signal detection pipeline) of the coalescence time of the signal of interest. This procedure allows to isolate one single peak at a time, and works well in the configurations we explored. However, as the time shift approaches zero, isolating one single peak at a time is not possible, and within our framework we can recover only the parameters for the louder signal (i.e., the highest peak in the posterior). In our approach, we are implicitly assuming that signal detection will return the coalescence times of the two signals with an uncertainty lower then $\mathcal{O}(10-100)$ ms. This is a reasonable assumption, which, however, needs to be tested by a dedicated analysis dealing with BBH signals' recovery in the context of 3G detectors (see also (Regimbau et al. 2012; Meacher et al. 2016) for the BNS case).

We also studied the emergence of biases in the overlapping signals scenarios considered, by varying some key parameters of the two signals such as their coalescence time, coalescence phase, and luminosity distance. We found that significant biases (at the $2-3\sigma$ level) arise in the range $-1\,\mathrm{s} < \tau < 0\,\mathrm{s}$, and that these biases are caused primarily by the relative phase of the two signals and only marginally by the relative difference of the SNRs. As suggested by our Fisher analysis (Fig. 6.2, upper left panel), these biases tend to become minor for $\tau < -1.5\,\mathrm{s}$ and $\tau > 0\,\mathrm{s}$.

Dealing with these biases needs a different approach that we did not attempt in this work. One possible solution is to broaden the parameter space searching for multiple signals in the same Bayesian inference run. This is the approach that previous works have shown to be feasible in the context of LISA data analysis (e.g., see (Cornish & Crowder 2005; Littenberg et al. 2020)). Such approach could significantly increase the computational costs of the Bayesian algorithms; however, this is compensated by the fact that - as suggested here - novel algorithms may be needed only for closely-coalescing signals, that are a very small minority of the total number of signals expected in 3G detectors. Using current estimates for the BBH rates in future detectors, we find that signals coalescing within 1 s are expected to be at most hundreds per year.

Another possible solution to the biases would be to create an iterative procedure where one hierarchically determines the parameters of louder signals (as inferred from search algorithms) and subtracts them from the data before analysing weaker ones (Cutler & Harms 2006; Sachdev et al. 2020). Currently, it is unclear which approach will perform better in the

context of 3G detectors, and further work is needed to gauge the potential of both approaches.

In our exploratory study we did not deal with the consequences of varying the mass parameters of the two signals, nor did we include in our analysis other source parameters such as companion spins, the position of the source in the sky and the orientation of the binary relative to the detector frame. The SNR range explored in our study (20-100; see also Tab. 6.1) is also limited compared to the range expected to be covered by 3G detectors (Punturo et al. 2010; Reitze et al. 2019b; Borhanian & Sathyaprakash 2022). In particular, when overlapping signals arrive from different positions in the sky then they would, in general, have different coalescence times in different detectors, which might help to isolate one of the peaks better (Christian et al. 2018). The inclusion of spins, on the other hand, introduces new physics in the formation of these overlapping signals such as spin precession, and may introduce another layer of complexity in the parameter inference problem (Fairhurst et al. 2020). These and related problems will be explored in a future study.

Acknowledgments

We thank Anuradha Samajdar, Justin Janquart, Chris Van Den Broeck and Tim Dietrich for sharing and discussing their results on a similar study (Samajdar et al. 2021). We are indebted to useful comments by Rory Smith and Salvatore Vitale. EP is grateful to Walter Del Pozzo for helpful suggestions. EP was supported by INFN in the framework of the 2019 NSF-INFN Summer exchange program. SS is supported by the Eberly Postdoctoral Fellowship of Penn State. BSS is supported in part by NSF Grant No. PHY-1836779, PHY-2012083, and AST-2006384. The authors are grateful for computational resources provided by the LIGO-Caltech Computing Cluster. This paper has the LIGO document number P2100044.

7 Constraining turbulence in protoplanetary discs using the gap contrast: an application to the DSHARP sample

Abstract

Constraining the strength of gas turbulence in protoplanetary discs is an open problem that has relevant implications for the physics of gas accretion and planet formation. In this work, we gauge the amount of turbulence in 6 of the discs observed in the DSHARP programme by indirectly measuring the vertical distribution of their dust component. We employ the differences in the gap contrasts observed along the major and the minor axes due to projection effects, and build a radiative transfer model to reproduce these features for different values of the dust scale heights. We find that (a) the scale heights that yield a better agreement with data are generally low ($\lesssim 4\,\mathrm{AU}$ at a radial distance of 100 AU), and in almost all cases we are only able to place upper limits on their exact values; these conclusions imply (assuming an average Stokes number of $\approx 10^{-2}$) low turbulence levels of $\alpha_{\mathrm{SS}} \lesssim 10^{-3} - 10^{-4}$; (b) for the 9 other systems we considered out of the DSHARP sample, our method yields no significant constraints on the disc vertical structure; we conclude that this is because these discs have either a low inclination or gaps that are not deep enough. Based on our analysis we provide an empirical criterion to assess whether a given disc is suitable to measure the vertical scale height.

Published in: **EP**, Giovanni P Rosotti, Benoît Tabone, *Constraining turbulence in protoplanetary discs using the gap contrast: an application to the DSHARP sample*, Monthly Notices of the Royal Astronomical Society, Volume 524, Issue 2, September 2023, Pages 3184–3200, doi.org/10.1093/mnras/stad2057
Reprinted here in its entirety.

7.1 Introduction

Characterising the magnitude of turbulence in accretion discs is a classical problem in astrophysics. This is because turbulence is often invoked (see historical discussion in Pringle 1981) as the mechanism responsible for powering accretion. On the one hand, therefore, the first scientific question that any such study seeks to address is whether the level of turbulence, commonly quantified through the $\alpha_{\rm SS}$ parameter (Shakura & Sunyaev 1973), is high enough to explain the observed accretion rates. For the specific case of proto-planetary discs we study in this paper, this is a particularly important question: the cold conditions of these discs, which are clearly in the non-ideal magnetohydrodynamics regime, make it far from obvious to understand whether the magneto-rotational instability (Balbus & Hawley 1991) can be a mechanism responsible for generating the required level of turbulence. Addressing this question, and studying in parallel other processes that could generate turbulence in proto-planetary discs, is a subject of many studies (see Lesur et al. 2022 for a recent review).

For proto-planetary discs, the issue runs even deeper than the question about accretion; even if turbulence was ultimately found not to be responsible for accretion, it would still affect a wealth of processes and therefore have a strong impact on planet formation. A non-exhaustive list of processes affected by turbulence includes the heating and cooling balance in the terrestrial planet-forming region due to the importance of viscous heating (Min et al. 2011), the diffusion of molecular species radially (Owen 2014) and vertically (Semenov & Wiebe 2011; Krijt et al. 2020), the diffusion of dust, setting both the dust disc vertical extent (Dubrulle et al. 1995) and the leakiness of dust traps (e.g., Zormpas et al. 2022). For what concerns planets in particular, turbulence has a profound impact on disc-planet interaction; its magnitude affects (Paardekooper et al. 2022) the ability of planets to open gaps in the disc and how fast they migrate by exchanging angular momentum with the disc. Turbulence is also a crucial parameter setting how quickly planets accrete gas (Bodenheimer et al. 2013) and dust (Johansen & Lambrechts 2017) from the disc, determining the final masses of planetary systems. Last but not least, turbulence controls the onset of the streaming instability (Drazkowska et al. 2022), one of the best known mechanisms for creating planetesimals and kick-starting the planet formation process.

It would thus be beneficial to have a method to constrain turbulence observationally. Thankfully, in the last few years, the field has been completely transformed by the Atacama Large Millimeter Array (ALMA), which provided order-of-magnitude improvements in sensitivity and angular resolution. First of all, by studying line broadening of emission lines, ALMA has allowed to directly detect turbulence in two discs (Flaherty et al. 2020), and only yielding upper limits in a limited number of other cases (Pinte et al.

2022). In addition, ALMA has opened up many other observational routes (recently reviewed in Rosotti 2023) for indirectly constraining turbulence. These routes include the study of the disc vertical thickness, the radial extent of dust and gas rings, and population studies that use disc demographics studies (see Manara et al. 2022 for a review), that is, catalogues of the fundamental disc properties (such as mass, radius and mass accretion rate) for large disc samples. In this way, in the last few years the study of disc turbulence has moved from an almost theoretical subject to an observational one.

In this paper, we constrain turbulence by measuring the disc vertical thickness. The vertical equilibrium of dust grains is a competition between settling, which is determined by the joint action of gas drag and gravity, and turbulence, which stirs up the grains in the vertical direction. In simple terms, then, the more turbulent the disc is, the thicker it is, but it should be highlighted that the presence of drag implies that the aerodynamic coupling between gas and dust (normally parametrized by the Stokes number St) also influences the thickness. Indeed, as we will recap in Sec. 7.5, the method is only sensitive to $\alpha_{\rm SS}/{\rm St}$.

We apply the technique developed by Pinte et al. (2016) in their study of HL Tau. The technique relies on the fact that many observed discs (Bae et al. 2022) present an emission pattern characterised by bright rings and dark gaps. Pinte et al. (2016) realised that due to projection effects in a disc with finite thickness the line of sight will intercept sections of the disc that are out of the midplane. In a gap, this has the effect that the adjacent bright regions partially contaminate the dark gap, lowering the gap depth. We will refer to this in the rest of the paper as the *gap-filling effect*. It is easy to realise that the geometry of projection is such that this filling effect is much larger along the minor axis of the disc than along the major axis. Once the image is deprojected in polar coordinates, as commonly done in the field, the resulting effect is that the gaps are more "filled" (i.e., shallower) along the disc minor axis and more "empty" (i.e., deeper) along the disc major axis. The difference between minor and major axes increases with the disc thickness and therefore it is a way to probe the vertical structure of the disc. A simple sketch in Figure 7.1 shows the simple geometrical argument behind the *gap-filling effect*. Extracting quantitative measurements from this effect requires building radiative transfer models of the emission.

So far, in addition to HL Tau, the method has been applied to HD163296 (Liu et al. 2022) and to Oph163131 (Villenave et al. 2022). The goal of this paper is to significantly expand this observational sample. For this purpose we selected the sample of the Large Programme DSHARP (Andrews et al. 2018), which consists of twenty discs imaged at 0.05" resolution, since it constitutes the largest homogeneous high-resolution survey of proto-planetary discs. We aim to determine in which cases this technique is successful in

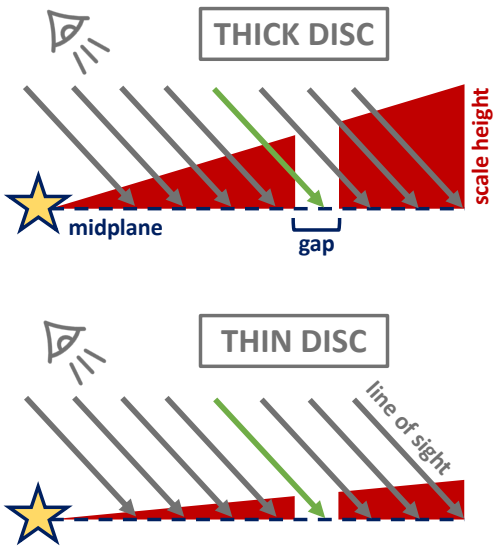

Figure 7.1: Sketch of the *gap-filling effect*. The lines of sight intercept the disc's plane with an angle that depends on the disc inclination. In the presence of a gap, lines of sights piercing through the gap (e.g., the one highlighted in green in the sketch) may still intercept sections of the disc that are far from the mid-plane due to simple geometrical effects. Therefore, the gap will be seen as partially filled by the observer. If the disc is thicker (thinner), this filling effect is stronger (weaker); this implies that the *gap-filling effect* can be used to indirectly gauge the vertical extension of the disc.

gauging the disc thickness, and, whenever possible, to place meaningful constraints on the disc scale heights.

This paper is structured as follows: in Sec. 7.2, we discuss the basic assumptions of our model and describe the steps of the analysis we perform to match DSHARP data. A description of the data sample is presented in Sec. 7.3. Sec. 7.4 presents the main results of the analysis, while Sec. 7.5 contains a discussion on the implications of our findings. Conclusions are given in Sec. 7.6.

7.2 Methods

In this section, we describe the basic assumptions of the model and provide details on the methodology we employ to compare our synthetic images with DSHARP observations.

7.2.1 Disc structure

In the following, we are only interested in modelling the dust component of the protoplanetary discs, as the ALMA observations we use are only focusing on the dust continuum emission. Therefore, in this section – whenever not stated otherwise – we refer with the term "disc" to the dust component only. In Sec. 7.5, we will discuss further how our results can be employed to study the relationship between dust and gas in the disc, and ultimately to constrain the amount of gas turbulence.

We model the disc as a cylindrically-symmetric system with a minimum radius $r_{\rm in}$ and a maximum radius $r_{\rm out}$. We assume for the vertical density distribution a Gaussian profile of the form:

$$\rho_d(r,z) = \frac{\Sigma(r)}{\sqrt{2\pi}h_d(r)} \exp\left(-\frac{z^2}{2h_d(r)^2}\right), \tag{7.1}$$

where $\Sigma(r)$ is the dust surface density and $h_d(r)$ is the dust scale height.

This profile originates from an analogy to the gas component, which can be assumed to be in hydrostatic equilibrium along the vertical direction and thus follows a Gaussian profile identical to eq. 7.1. Formally, the dust has a different equilibrium solution, but eq. 7.1 is a close approximation (Dubrulle et al. 1995). Furthermore, defining the dust density in this way is convenient as we can consider the ratio between gas and dust scale-heights, which will be important to estimate turbulence.

In Sec. 7.2.2, we discuss our procedure to determine the disc surface density $\Sigma(r)$. As for the dust scale height profile, we assume a simple flaring model:

$$h_d(r) = H_d \left(\frac{r}{r_0}\right)^{1.25}, \tag{7.2}$$

In what follows, we set the reference radius r_0 to 100 AU and take the value of the dust scale height at this radius, $H_d = h_d(r = 100\,\mathrm{AU})$, as the only free parameter of our model. The goal of our work is to gauge the value of H_d using the *gap-filling effect* on the disc minor axis.

In order to predict the observed surface brightness of the disc, we need to determine the dust temperature profile. For this, we assume that the dust is passively heated by the central star, and that a fraction $\phi_{\rm flux}$ of the total flux emitted by the star is intercepted by the disc. Following radiative-transfer models (e.g., Chiang & Goldreich 1997; D'Alessio et al. 1998; Dullemond et al. 2001, 2018), we set $\phi_{\rm flux} = 0.02$ and write (σ is the Stefan-Boltzmann constant):

$$T(r) = \left(\frac{0.5\phi_{\rm flux}L_\star}{4\pi r^2\sigma}\right)^{1/4} = T_{\rm in}\left(\frac{r_{\rm in}}{r}\right)^{0.5}, \tag{7.3}$$

where we have expressed everything in terms of the temperature at the inner radius, $T_{\rm in}$. For the sake of simplicity, we use this analytical description of

the disc temperature for our model – instead of self-consistently computing the temperature using a Montecarlo approach (see e.g., Liu et al. 2022). In Sec. 7.5.4, we discuss the reasons for such choice and the caveats that come with it.

The expected surface brightness of the disc can then be determined assuming dust thermal emission. In order to create mock images of our disc models, we use the code RADMC-3D[1]. We set the extrinsic parameters (such as the distance, sky coordinates, inclination, position angle) in accordance with observations (see Sec. 7.3), and we produce synthetic images of the discs according to the radiative transfer computation. Then, we use the CASA package (CASA Team et al. 2022) to produce mock observations with the same beam and antennae configuration of the original ALMA data. In order to do this we have retrieved from the DSHARP Data Release webpage[2] the visibility files of the DSHARP observations. In our analysis we use the same version of CASA used by the DSHARP team (v 5.1.1-5) to ensure that the data and the models have been processed in the same way[3]. We created synthetic visibilities from the radiative transfer image at the uv coordinates of the observations using the CASA task ft. We then apply the CLEAN algorithm to generate a synthetic ALMA image to compare with the observed image. We use the scripts provided by the DSHARP team in order to make sure that we use the same CLEAN parameters as the observations. To reduce the computational time, it is common in the field to employ the simpler approach of a convolution with a Gaussian beam. While this is often satisfactory, we noticed in early tests that the detailed shape of the emission profile in the gap is different from images produced by the CLEAN algorithm. In addition, some of the DSHARP sources have clear CLEAN artefacts such as negative emission that cannot be reproduced with a simple Gaussian convolution. Therefore, we adopt here a consistent approach to include the contribution of these cleaning artifacts.

7.2.2 Inferring the disc surface density

In order to proceed further with our analysis, we need to infer the surface density $\Sigma(r)$ of the observed disc. This is not straightforward, as simple power-law models are not capable to reproduce the wealth of substructures (gaps and rings) that are observed in the DSHARP images. Given that our goal is to use the *gap-filling effect* as a probe of the disc vertical size, modeling these substructures within a reliable framework is of paramount importance. Therefore, similarly to what was done in Pinte et al. (2016), we

[1] http://www.ita.uni-heidelberg.de/~dullemond/software/radmc-3d/
[2] https://almascience.eso.org/almadata/lp/DSHARP/
[3] That being said, for safety we have recomputed the CLEAN images for the data starting from the visibility files, in order to be sure that we use the same CLEAN parameters in the data as in the models.

employ here an iterative procedure to find the correct surface density of our discs. We outline the procedure in the following paragraphs, and provide an overview of the different steps involved in the iteration cycle in Figure 7.2.

The fundamental idea we adopt in this procedure is that the intensity observed along the major axis is a good proxy for the real surface density of the disc. This is because, as already discussed in Sec. 7.1, the *gap-filling effect* affects only marginally the major axis, whereas it has the strongest effect on the minor axis. Therefore, our goal is to find via multiple iterations a surface density profile that is able to match the intensity observed along the major axis.

The procedure can be summarized as follows. First of all, we need to extract the intensity along the major axis, $I_{\rm maj}^{\rm (data)}(r)$, from the 2D images. In order to do that, for every disc we analyze, deproject the image and average two opposite slices of 1/8 (i.e., with a width of $\pi/4$) of the disc centered on the major axis. When deprojecting the disc emission maps, we make sure that the images are aligned with the discs' centres by using the offsets in the x and y coordinates reported by Huang et al. (2018) (see their Tab. 2). The resulting $I_{\rm maj}^{\rm (data)}(r)$ represents our benchmark profile that we aim to reproduce with a suitable choice of the disc surface density.

Then, we use as a first guess for the surface density profile the output of the FRANKENSTEIN (Jennings et al. 2020) fit of the DSHARP sources presented in Jennings et al. (2022). FRANKENSTEIN is a code that uses Gaussian processing to fit disc emission profiles in visibility space, using the assumption that the emission is azimuthally symmetric. This gives a good starting point for the initial surface density since FRANKENSTEIN can achieve a spatial resolution higher than the CLEANed images we analyse in this paper. While this gives us the shape of the surface density, note that we also need a normalisation constant: Frankestein fits for the emission profile (giving a profile $I_\nu^{\rm (FRANK)}(r)$ as an output), while we need a surface density to give as input to RADMC-3D. In order to convert the intensity profile into a surface density, we use as a constraint the formula often employed (Beckwith et al. 1990) to estimate the disc mass $M_{\rm dust}$ from sub-mm observations:

$$M_{\rm dust} = \frac{F_\nu d^2}{\kappa_\nu B_\nu(T_{\rm dust})} \tag{7.4}$$

where F_ν is the flux in the image, d the distance to the source; for $T_{\rm dust}$ we take a temperature of 20 K and κ_ν is the opacity of the dust we employ. Since we consider a single grain population, the physical quantity we are constraining is the dust optical depth (given the prescribed temperature profile), and not the dust surface density. This implies that the value of the opacity only acts as a normalisation for the dust surface density and does not have any influence on our conclusions - with a different dust opacity we would simply need to change the dust surface density accordingly in order

to have the same optical depth. Notice also that the formula is only an approximation (the emission is not guaranteed to be optically thin and 20 K may not be the correct value); however the value reported above is only needed to kickstart the iteration and the iterative procedure will take care of reaching the correct values, both for the normalisation and for the shape of the surface density. Finally, because we fitted for the emission profile but need the surface density, we multiply the resulting profile by $r^{1/2}$ to take into account the variation in disc temperature with radius[4]. Let us call $\Sigma_0^{(\mathrm{guess})}(r)$ the surface density we have obtained in this way. We use this guess to define our disc structure (setting also the dust scale height parameter H_d to a fixed value) and produce synthetic observations using RADMC3D + CASA (see Sec. 7.2.1).

Subsequently, we apply the same procedure as described above to deproject these mock observations and to extract a mock intensity profile along the major axis[5], $I_{\mathrm{maj},0}^{(\mathrm{guess})}(r)$. This profile can be directly compared to the observational one, $I_{\mathrm{maj}}^{(\mathrm{data})}(r)$. This comparison outputs a ratio, $\xi_0(r) = I_{\mathrm{maj}}^{(\mathrm{data})}/I_{\mathrm{maj},0}^{(\mathrm{guess})}$ that parametrizes how well the initial guess for the surface density is able to reproduce observations. We can improve this match simply by multiplying the initial guess for the surface density profile, $\Sigma_0^{(\mathrm{guess})}(r)$, for the ratio $\xi_0(r)$, finding a new guess for the disc surface density $\Sigma_1^{(\mathrm{guess})}(r)$. To prevent large variations of the surface density from one iteration to the next, we do not allow variations larger than a factor 4 in a single iteration. We then iterate this procedure by using this new surface density profile to produce mock observations $I_{\mathrm{maj},1}^{(\mathrm{guess})}$ and update the surface density using the sequence:

$$\Sigma_{n+1}^{(\mathrm{guess})}(r) = \xi_n(r)\Sigma_n^{(\mathrm{guess})}(r) = \frac{I_{\mathrm{maj}}^{(\mathrm{data})}}{I_{\mathrm{maj,n}}^{(\mathrm{guess})}}\Sigma_n^{(\mathrm{guess})}(r) \tag{7.5}$$

We stop the iteration when a value of $|\xi(r) - 1| < 0.05$ is reached for every radius r. On average, this takes around $10 - 15$ cycles. As expected, the convergence is very easily achieved where the intensity profile is smooth, whereas it takes more iterations in the regions where gaps and rings are present, especially when they are narrow and deep. For a few systems, this implies that convergence is not reached at the bottom of the deepest gaps even after 15 iterations, with the difference between the model and data being in the range $5-10\%$. We empirically find that increasing the number of iterations does not give any significant advantage for these peculiar systems, with only minimal gains in terms of model-data accordance despite the large

[4] In the same way as for the normalisation, this is only a first order correction; the iterative procedure will better refine this radial scaling.

[5] With the exception of AS 209, where the procedure described here uses the azimuthally averaged intensity profile rather than the major axis (see also Tab. 7.1).

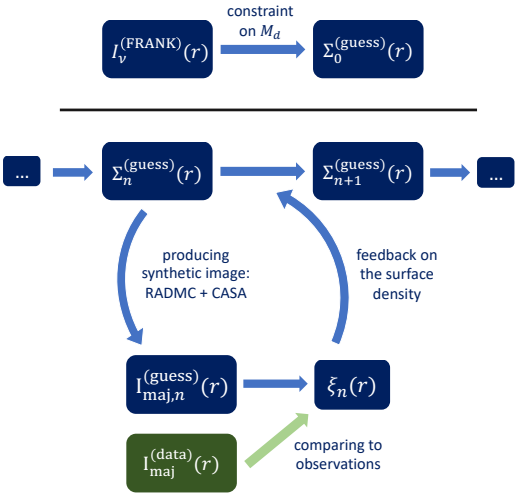

Figure 7.2: Overview of the iteration procedure we employ to extract the disc surface density, $\Sigma(r)$. As described in Sec. 7.2.2, we choose the profile for the surface density that matches the observed intenstity profile along the major axis. The various quantities appearing in the sketch are defined in the main text. The top part of the sketch refers to the kickstarting of the process (where we find a first guess for the surface density profile), while the bottom part shows an instance of a single iteration.

number of iterations employed. Therefore, we decide to set a maximum iteration number $n = 15$ and insert a caveat for the systems that are not converged inside the gaps according to our criteria (Sec. 7.3).

7.2.3 Analysis of the gaps filling effect on the minor axis

At the end of the iteration cycle described in Sec. 7.2.2, we obtain, for a fixed value of dust scale height H_d, a fiducial profile for the disc surface density, $\Sigma^{(\mathrm{model})}(r; H_d)$. Note that the dependence of $\Sigma^{(\mathrm{model})}(r; H_d)$ on H_d is very mild, as the surface density is obtained by comparing the model with the data along the major axis, where the vertical structure of the disc has only a small effect on the final intensity. This $\Sigma^{(\mathrm{model})}(r; H_d)$ corresponds to an intensity profile along the major axis – $I_{\mathrm{maj}}^{(\mathrm{model})}(r; H_d)$ – that matches the observed one – $I_{\mathrm{maj}}^{(\mathrm{data})}(r)$ (see also the sketch in Fig. 7.2).

Our goal is then to extract, both from the observational data and from our fiducial model, the intensity profiles along the minor axis – $I_{\mathrm{min}}^{(\mathrm{data})}(r)$ and $I_{\mathrm{min}}^{(\mathrm{model})}(r; H_d)$, respectively. Similarly to what described in Sec. 7.2.2,

in order to do this, we take the deprojected 2D images and average two opposite slices of 1/8 of the disc centered on the minor axis.

Along the minor axis, the predicted intensity can depend quite strongly on the value of the dust scale height H_d, as the vertical thickness of the disc directly influences the gap filling along the minor axis. Therefore, a simple comparison between $I_{\rm min}^{\rm (data)}(r)$ and $I_{\rm min}^{\rm (model)}(r; H_d)$ offers a way to place constraints on the vertical structure of the disc. A quantitative analysis of this comparison and the implications of the results we find are presented in Sec. 7.4 and 7.5.

7.3 DSHARP data sample

In this section, we describe the systems we use for our analysis. DSHARP is a very high resolution (∼0.035", or 5 AU) observational campaign that targeted 20 proto-planetary discs with the goal of finding and characterising substructures in the dust continuum emission at 240 GHz. We examined the entire DSHARP catalogue and excluded the systems that are not suitable for our study of the *gap-filling effect*. These include 3 single systems that show signs of spirals (i.e, IM Lup, Elias 27, and WaOph 6) and two binary systems (HT Lup and AS 205), where the individual discs either show signs of spirals or lack clear substructure. Spirals are not compatible with the assumption of perfect azimuthal symmetry in our disc model and we therefore discard the discs showing this signature.

For the remaining 15 systems, we run our model to find the best matching value(s) of the dust scale height H_d. We describe the results of this analysis in the following section. Here, we provide more details on the properties of these systems. In Table 7.1, we report the parameters of the systems as listed by Andrews et al. (2018): these include the mass and luminosity of the central star, the distance of the system, the inclination angle, the position angle (PA), the outer radius of the disc, and the beam size. The inner radius cannot be determined easily from observations, and thus we always set it to $r_{\rm in} = 2\,{\rm AU}$. This choice has no relevant impact on the final results since we are only interested here in radial locations with gaps.

7.4 Results

In this section, we apply the analysis described in Sec. 7.2 to the data sample presented in Sec. 7.3. In order to follow in detail the different steps of our analysis, we first focus on a single instance (i.e., GW Lup), and then we provide an overview of the global results for the rest of the sample we considered.

Table 7.1: Properties of the DSHARP systems considered here for the analysis. Observational parameters are taken from Andrews et al. (2018) and Huang et al. (2018). From left to right: name of the source; mass and luminosity of the central star ($M_\star$, $L_\star$), distance (D), inclination angle (i), position angle (PA), maximum radius ($r_{\rm out}$), beam size (beam); results from our analysis, i.e., the best-fitting value of H_d and the associated reduced chi-squared, $\chi^2_{\rm norm}$. The final column describes whether it was possible to use our analysis to constrain the scale height in a meaningful range. The units of measurements for every parameters are shown on the second row; uncertainties for the disc extensions and the beam size are irrelevant to our discussion and not shown here.

	Observational data							Model Predictions		
ID	$\log M_\star$	$\log L_\star$	D	i	PA	$r_{\rm out}$	beam	H_d	$\chi^2_{\rm norm}$	Constraint
	[$M_\odot$]	[$L_\odot$]	[pc]	[°]	[°]	[AU]	[" × "]	[AU]		on H_d
GW Lup	$-0.34^{+0.10}_{-0.17}$	-0.48 ± 0.20	155 ± 3	38.7 ± 0.3	37.6 ± 0.5	105	0.045×0.043	$\lesssim 4$	0.27	Upper limit
DoAr 25	$-0.02^{+0.04}_{-0.19}$	-0.02 ± 0.20	138 ± 3	67.4 ± 0.2	110.6 ± 0.2	165	0.041×0.022	$\lesssim 2$	4.5	Upper limit
Elias 24 [a]	$-0.11^{+0.16}_{-0.08}$	0.78 ± 0.20	136 ± 3	29.0 ± 0.3	45.7 ± 0.7	160	0.037×0.034	$\lesssim 2$	4.9	Upper limit
HD 142666	$0.20^{+0.04}_{-0.01}$	0.96 ± 0.21	148 ± 2	62.22 ± 0.14	162.11 ± 0.15	80	0.032×0.022	–	–	No constraint
AS 209 [c]	$-0.08^{+0.11}_{-0.14}$	0.15 ± 0.20	121 ± 2	34.97 ± 0.13	85.76 ± 0.16	160	0.038×0.036	$\lesssim 2$	0.79	Upper limit
Elias 20	$-0.32^{+0.12}_{-0.07}$	0.35 ± 0.20	138 ± 5	49 ± 1	153.2 ± 1.3	85	0.032×0.023	–	–	No constraint
Sz 129	$-0.08^{+0.03}_{-0.15}$	-0.36 ± 0.20	161 ± 3	34.1 ± 1.3	151 ± 2	95	0.044×0.031	–	–	No constraint
HD 163296 [b]	$0.31^{+0.05}_{-0.03}$	1.23 ± 0.30	101 ± 2	46.7 ± 0.1	133.33 ± 0.15	170	0.048×0.038	≈ 4	25	Upper/lower lim.
HD 143006	$0.25^{+0.05}_{-0.08}$	0.58 ± 0.15	165 ± 5	18.6 ± 0.8	169 ± 2	105	0.046×0.045	–	–	No constraint
SR 4 [a]	$-0.17^{+0.11}_{-0.14}$	0.07 ± 0.20	134 ± 2	22 ± 2	18 ± 5	85	0.034×0.034	–	–	No constraint
RU Lup	$-0.20^{+0.12}_{-0.11}$	0.16 ± 0.20	159 ± 3	18.8 ± 1.6	121 ± 5	80	0.025×0.024	–	–	No constraint
MY Lup	$0.09^{+0.03}_{-0.13}$	-0.06 ± 0.20	156 ± 3	73.2 ± 0.1	58.8 ± 0.1	115	0.044×0.043	$\lesssim 4$	0.58	Upper limit
Sz 114	$-0.76^{+0.08}_{-0.07}$	-0.69 ± 0.20	162 ± 3	21.3 ± 1.3	165 ± 4	65	0.067×0.028	–	–	No constraint
WSB 52	$-0.32^{+0.13}_{-0.17}$	-0.15 ± 0.20	136 ± 3	54.4 ± 0.3	138.4 ± 0.3	39	0.033×0.027	–	–	No constraint
DoAr 33	$0.04^{+0.05}_{-0.17}$	-0.18 ± 0.20	139 ± 2	41.8 ± 0.8	81.1 ± 1.2	27	0.037×0.024	–	–	No constraint

a Convergence (described in Sec. 7.2.2) cannot be reached. The (maximum) relative difference between the model and the data does not go below the threshold value of 5%; instead, it sits between 5% and 10%.

b When determining the intensity along the major axis, we use only one side of the major axis since the other one presents a feature that disrupts azimuthal symmetry.

c For this system, in order to determine the surface density via our convergence procedure, we do not consider the intensity along the major axis but the average. This is because the former presents some negative values in the gaps and this creates an issue with the convergence procedure described in Sec. 7.2.2.

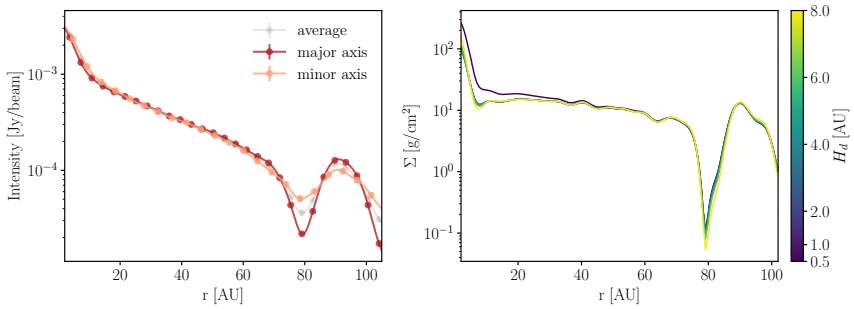

Figure 7.3: *Left:* Intensity profiles of GW Lup along the minor ($I^{(\mathrm{data})}_{\mathrm{min}}$; salmon line) and major ($I^{(\mathrm{data})}_{\mathrm{maj}}$; brick) axes, together with the azimuthally averaged profile (gray line). Profiles along the two axes are extracted according to the procedure described in Sec. 7.2.2. Errors are computed according to what outlined in Sec. 7.4.1. *Right:* Surface density profiles $\Sigma^{(\mathrm{model})}(r; H_d)$ predicted by our model for different disc scale heights H_d. The profiles are obtained by using the iteration procedure described in Sec. 7.2.2.

7.4.1 GW Lup as a case study

GW Lup is a disc with an average inclination of $i \approx 40°$ and a major structure composed by a gap at $r \approx 74$ AU and a ring at $r \approx 85$ AU (Huang et al. 2018). Because of these properties, it is very well suited for an exemplification of our methodology.

In the first step of our analysis, we take the observational image and extract the profiles along the major and the minor axes as described in Sec. 7.2 (i.e., by averaging the deprojected images in slices that are centred on the axes and have an angular size of $\pi/4$). The resulting profiles are shown in Fig. 7.3 (left panel). We also show the profile obtained by averaging the deprojected image in concentric rings (i.e., the azimuthally-averaged profile). All the different profiles clearly show the characteristic structure of the gap + ring feature. However, as expected, the intensity contrast along the minor axes is smaller due to the *gap-filling effect*.

In order to quantify the statistical uncertainty on the three profiles, we simply compute the standard deviation σ of the data in the deprojected images along the slices (or rings) considered, and divide it by the square root of the number of independent data points considered ($\sqrt{N_{\mathrm{beams}}}$). This latter quantity is simply the azimuthal extent of the slice/ring $\Delta\phi R$, divided by the size of the beam – which we obtain by averaging the two axes of the beam; along the minor axis, we increase the size of the beam by a factor $\cos i$ to take projection into account. In formula, we get: $\sigma_{\mathrm{profile}} = \sigma\sqrt{\mathrm{beam}/\Delta\phi R}$. In Fig. 7.3 (left panel), we plot the error bars only every N_{beams}, so that they are independent of each other. Note that these errors are very small, and therefore hardly visible in the scale of the plot.

As a second step, we choose a value for the dust scale height parameter H_d. In what follows (where not stated otherwise), we consider the following set of values for H_d: $\{0.5, 1, 2, 4, 6, 8\}$ AU. For each of these values, then, we apply an iteration procedure to match the intensity along the major axis, with the goal of finding the best surface density for the disc (see Sec. 7.2 for details on this iteration procedure). The right panel of Figure 7.3 shows the fiducial surface density output by our iteration cycle for different H_d. As expected, the predicted surface density is almost identical for different H_d values (with the notable exception of $H_d = 0.5$ AU).

Using these fiducial profiles for the surface density, we can produce mock observations setting the same observational parameters as in Tab. 7.1 and using the same configuration as the data (see Sec. 7.2.1 for more details on mock images generation). Figure 7.4 shows these mock images for the two extreme H_d values of $H_d = 0.5$ AU and $H_d = 8$ AU, together with the real observations from DSHARP. Even a quick look at the figures allows us to appreciate how the different systems have a similar intensity along the major axis, whereas they present a different gap filling along the minor one, with the image of the thick disc being significantly more blurred than the one referring to the thin disc.

This difference can be quantified by deprojecting the images and extracting the profiles along the major and the minor axes in the same way as done with the observational data (i.e., averaging two 1/8-slices of the deprojected images centered along the axes). The resulting profiles for the major (minor) axis are shown in the left (right) panel of Fig. 7.5, together with the same observational data that are also shown in Fig. 7.3 (left panel). Given that we are interested in the *gap-filling effect*, in the following, we focus only on the region where the gap+ring structure resides (i.e., between 70 and 95 AU).

As expected, the intensity along the major axis is almost the same for any values of the disc scale height H_d: all of the different profiles are perfectly compatible with the data. The azimuthally-average intensity from observations is also shown as a reference, in order to highlight how the data vary along different azimuthal axes. The intensity along the minor axis (right panel), on the other hand, strongly varies with H_d. In this plot, we can appreciate the predictive power of our method: the *gap-filling effect* implies that for large values of the disc scale height $H_d \gtrsim 6$ AU the resulting profile is much smoother (i.e., the gap is much more filled) with respect to the thin disc cases ($H_d \lesssim 4$ AU). Given that the gap in the original data image (salmon data points) is considerably empty, we can conclude that the latter case is to be preferred by observations. Indeed, only the lines with $H_d \lesssim 4$ AU are compatible with the intensity profile of the gap + ring shown by the data. Therefore, we can conclude that the disc GW Lup is thinner than ≈ 4 AU at $r = 100$ AU. In the last columns of Tab. 7.1, we report this

conclusion by indicating the constraints we get on the scale height parameter H_d.

In order to quantify the agreement between observations and our mock profiles along the minor axis, we choose to employ the χ^2 statistics. However, we caveat that our aim is not to compare models and data in a way that sits on solid statistical bases. This is because, although our iterative procedure works quite well, discrepancies at the level of few percent from the observed emission remain (even along the semi-major axis). These discrepancies are significant given the signal to noise of the observations; in other words, the noise in the data is smaller than our ability to build radiative transfer models that reproduce them. This is a systematic source of error that is not accounted for in a statistics like the χ^2. This does not entail that our method is flawed: in practice, the difference brought upon by the *gap-filling effect* is much larger than the residual discrepancy between data and model. However, given the issues with a detailed comparison between our model and data, we note that the absolute value of χ^2 should not be used to accept or reject models, as it would be the case in a regular statistic test. Nevertheless, for completeness, we report the minimum value of the reduced chi-squared ($\chi^2_{\rm norm}$) in the second to last column of Table 7.1. This is the chi-squared divided by the number of degree of freedom (i.e., the number of independent data points + the number of free parameters in the model). We stress the fact that this number, however, does not have statistical validity and it is not a good parameter to accept/reject our model.

Instead, it is useful to employ the χ^2 as a way to test which of the values of H_d considered in the analysis has a better quantitative agreement with the data. In Fig. 7.7, we plot the logarithm of the likelihood function (i.e., $\log \mathcal{L} \propto -\chi^2/2$) normalized to its peak value, for different values of the parameter H_d. GW Lup is shown in blue, whereas all the other systems for which we get meaningful constraints on the scale height (see Sec. 7.4.2) are shown in the same plot with different colours.

From Fig. 7.7, we can confirm visually that the best fitting value of the disc scale height is $H_d = 4\,\mathrm{AU}$. However, values that are smaller than 4 AU are also compatible with data, as the value of the likelihood is smaller but still comparable, especially for $H_d = 1\,\mathrm{AU}$ and $H_d = 2\,\mathrm{AU}$. Values of H_d greater than 4 AU have significantly smaller likelihoods, and therefore are rejected by our analysis.

7.4.2 Overview of the other systems

In this section, we present the results of our analysis for the remaining systems considered in Sec. 7.3. In Tab. 7.1 (last columns), we show the constraints we are able to place on the values of the disc scale height based on the comparison between our model and the data along the minor axis. For most of the systems, however, we find that we are unable to place any

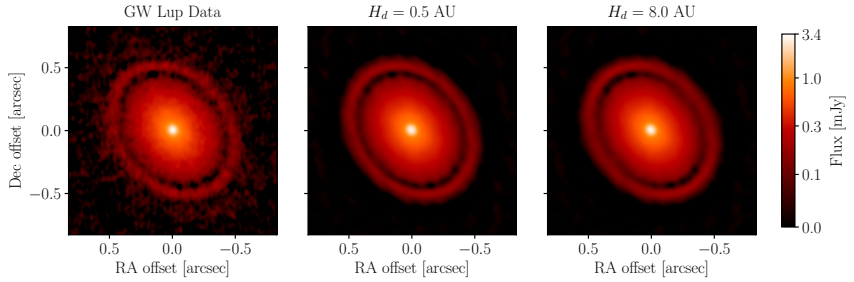

Figure 7.4: Original image of GW Lup (top left) from the DSHARP survey (Andrews et al. 2018), together with our mock images created using two extreme values of H_d ($H_d = 0.5$ AU and $H_d = 8$ AU) as well as the surface density profiles shown in Fig. 7.4. All images are plotted using an asinh stretch. Mock images are obtained using the same CLEAN settings as used for the data in DSHARP (Andrews et al. 2018).

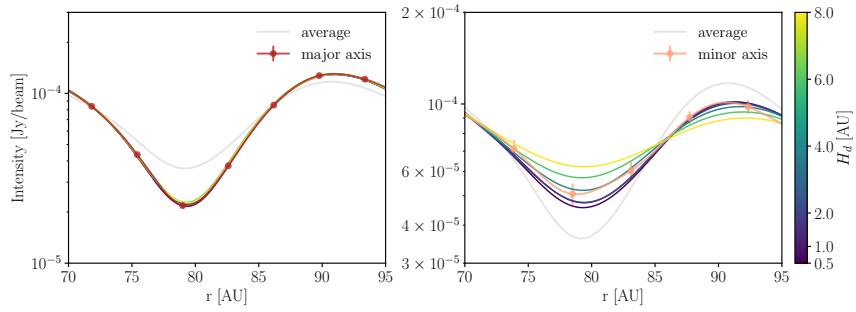

Figure 7.5: *Left:* Comparison of the intensity profiles for GW Lup predicted by our disc model along the major axis (for different values of the disc scale height H_d; see color bar) and the one extracted from data (firebrick points). For reference, the observed intensity averaged over the whole azimuthal angle is also plotted with a transparent grey line. The plot only shows a small section of the disc between $r = 70$ AU and $r = 95$ AU, where the major substructures (gap + ring) of GW Lup are present. Predicted intensities align very well with data points, and therefore are almost indistinguishable in the plot. *Right*: Same as the left panel, but for the predicted (see color bar) and observed (salmon data points) intensities along the minor axis. Different values of H_d are connected with very different predicted intensities, and this allows us to constrain the true value of H_d.

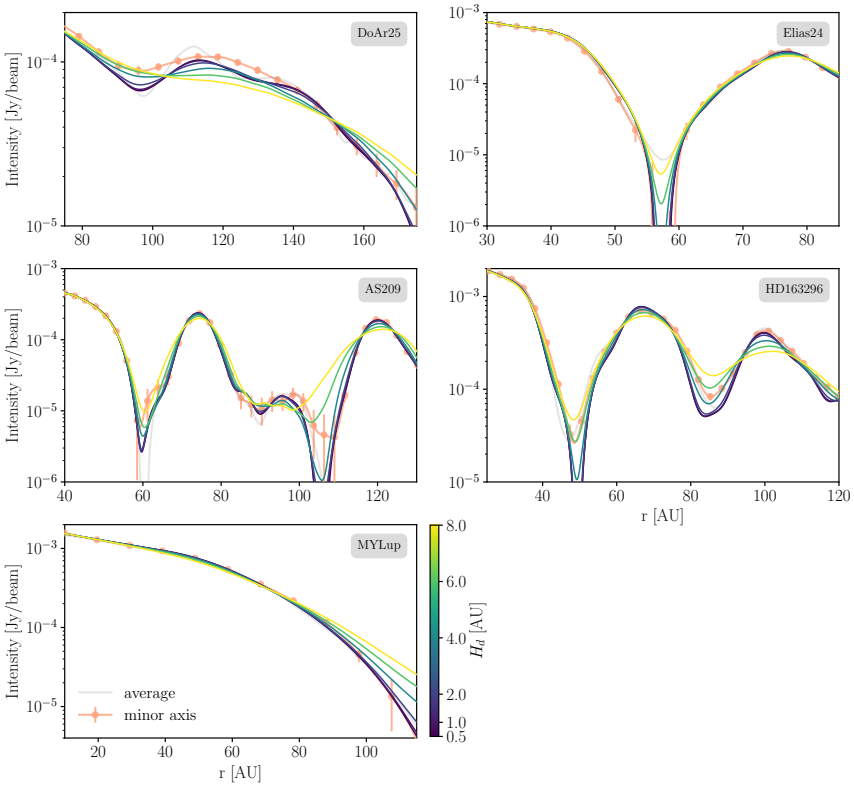

Figure 7.6: Same as the right panel of Fig. 7.5, but for the systems discussed in Sec. 7.4.2: DoAr 25 (top left), Elias 24 (top right), AS 209 (middle left), HD 163296 (middle right), MY Lup (bottom left).

constraints. This is because the *gap-filling effect* in those systems is not strong enough to produce significant effects on the final predicted intensities. Indeed, we find that for those systems different H_d values produce very similar profiles even along the minor axis. This implies that our approach is not effective in these cases given the structure of the emission and the resolution of observations. We discuss further in Sec. 7.5.3 under which conditions our method is effective in determining the discs' vertical structure.

We choose to focus our discussion in this section only on the systems that yield relevant constraints on H_d. The remaining systems – where our method fails to apply – are presented in Appendix 7.A.

Figure 7.6 shows the model intensities along the minor axes for the systems where our method is successful in gauging H_d, together with the observed intensities along the same axis (see also the right panel of Fig. 7.5 for the case of GW Lup). In all of these plots (and in the fitting routine), we focus only on the sections of the discs that are relevant for the application of our method and allow us to constrain the value of H_d (e.g., major gaps/rings).

Analogue plots are shown in Appendix 7.B (Fig. 7.10) for the same systems, but focusing on the major axis instead of the minor one – same as the left panel of Fig. 7.5, where we focus on the results of GW Lup only. Intensities along the major axis are generally well recovered by our model because the aim of the convergence procedure described in Sec. 7.2.2 is to correctly reproduce the observed intensity along this axis. Therefore, this figure serves as a reference in order to test the validity of our approach. For completeness, we also include mock images of these systems for $H_d = 0.5\,\mathrm{AU}$ and $H_d = 8\,\mathrm{AU}$ and compare them with observations in Appendix 7.B.

In Fig. 7.7, instead, we show the log-likelihood as a function of the parameter H_d (normalized to the peak value) for all the systems together. The log-likelihood is computed according to the models and data profiles that are shown in Figure 7.6 (i.e., the intensities along the minor axis).

In the following, we discuss the results of these figures for each system individually.

7.4.2.1 DoAr 25

Due to its large inclination angle ($i \approx 67°$) and to the presence of a major gap structure, DoAr 25 is a disc where the *gap-filling effect* is quite prominent. Therefore, we expect our method to be effective in discerning which disc scale height is compatible with the observed emission. Indeed, we see (Fig. 7.6) that different scale heights give rise to very different intensity profiles along the minor axes. However, none of these profiles is perfectly compatible with the observed emission. In fact, the observed gap + ring structure presents an offset with respect to all of the synthetic ones, making it hard to achieve a fair comparison between observations and models. The origin of this offset is unclear; we remark that all the models are converged and can

well reproduce the emission profile along the major axis, as can be seen in Fig. 7.10, at least out to 150 AU – beyond which the observations became relatively noisy. The offset may be due to an intrinsic asymmetry in the disc, whereas in our approach we had to assume that the disc is symmetric and any asymmetry is coming from radiative transfer and projection effects [6].

Nonetheless, we note that the presence of an observable gap in the minor axis' intensity profile is already a significant probe of a very small disc scale height. This is because, due to the high inclination of the disc, any values of H_d that are $\gtrsim 4$ AU would result in an almost complete filling of the gap. Therefore, we conclude that only $H_d \lesssim 2$ AU values are compatible with the observed gap + ring structure along the minor axis. This conclusion is supported by the χ^2 analysis (Fig. 7.7), in which we find that the best fitting value of the disc scale height is $H_d = 2$ AU, with small values significantly preferred to larger ones.

7.4.2.2 Elias 24

Elias 24 has a very wide and deep gap around $r \approx 57$ AU. The gap is so deep that, even along the minor axis, the intensity profile presents some negative values. These negatives are due to artefacts created by the CLEAN algorithm; it is notable that they are not present along the major axis and in the averaged profile (gray transparent line). However, given the fact that we adopt the same cleaning procedure as the one used for the data, we can correctly reproduce the profile even when it becomes negative.

Such a deep gap implies, once again, that the disc scale height is very small: only the profiles for $H_d \lesssim 2$ AU show an intensity that becomes negative in the gap centre, whereas larger values of H_d imply at least a partial gap filling along the minor axis and fail to reproduce the CLEANing artefacts. The best χ^2 value, as expected, sits in the range $H_d = 0.5 - 2$ AU.

7.4.2.3 AS 209

The intensity profile of AS 209 is particularly complex: many substructures can be identified both in the inner region of the disc and in the outer one (Huang et al. 2018). However, only three outer gaps are deep enough to be considered for our analysis of the filling effect. A first major gap is present at $r \approx 61$ AU, whereas two other gaps $r \approx 90$ AU and $r \approx 105$ AU form a large, single structure that is delimited by two bright rings at $r \approx 74$ AU and $r \approx 120$ AU, respectively. Therefore, in our analysis, we use this region ($40\,\mathrm{AU} < r < 130\,\mathrm{AU}$) to study how the predicted intensities compare with the data.

[6] We have also tried to vary the disc optical depth by changing the normalization of the temperature profile by a factor 2 in either direction, in order to investigate whether optical depth effects could be the cause of the offset. However, we found it to not be the case: results presented here are valid for all the models we experimented with.

Due to the complexity of the observed intensity profile, however, it is hard to tell which profile fits the data better by simply looking at Fig. 7.6. One thing that is particularly easy to observe is that gaps are fairly deep (and rings fairly bright), and thus very large values of H_d – represented by green/yellow lines – are to be excluded. See for example how the green line ($H_d = 6\,\mathrm{AU}$) fails significantly to reproduce the depth of the gap along the minor axis at 105 AU (Fig. 7.6), while being a good fit to the major axis (Fig. 7.10). The χ^2 analysis can quantify this, and it confirms that very small values of the disc scale height ($H \lesssim 2\,\mathrm{AU}$) are preferred over larger ones.

7.4.2.4 HD 163296

HD 163296 is another disc whose morphology is very promising for the application of our method. It has an inclination of $i \approx 47$ deg, and two major gaps at $r \approx 48\,\mathrm{AU}$ and $r \approx 86\,\mathrm{AU}$. The first gap is quite peculiar, as the emission map shows a sizeable blob in the gap along the major axis. This blob represents an issue for our disc modelling, as it is an obvious breaking of azimuthal symmetry. Therefore, we choose to exclude the region containing the blob from our analysis. In order to do that, whenever computing the intensity along the major axis (e.g., to find the surface density with the iteration procedure outlined in Sec. 7.2.2), we select only the slice on the side where the blob is not present. We double-check that this choice does not have an influence on the results we find for the outer gap by running a model that includes both sides of the major axis (therefore including the blob, so that the model is only meaningful for the outer gap) and confirming that we obtain very similar emission profiles along the minor axis for the outer gap region.

Looking at the intensity profiles along the minor axis (Fig. 7.6), we note that there is a broad agreement with data for values of the disc scale height in the range $1\,\mathrm{AU} < H_d < 6\,\mathrm{AU}$, depending on the exact gap/ring considered. As a rule of thumb, both gaps are well-fitted by relatively large scale heights ($H_d \approx 4 - 6\,\mathrm{AU}$), whereas the two rings seem to be compatible with lower values of H_d. The overall agreement is captured by our χ^2 analysis, which reveals a very strong preference for an intermediate value of the disc scale height ($H_d = 4\,\mathrm{AU}$). Therefore, this disc is the only one for which we can place relatively solid constraints on both the upper and the lower limits of the disc thickness. We caveat the reader, however, that the strength of this constraint should be not overestimated. In fact, as also discussed in Sec. 7.4.1, our χ^2 analysis does not take into account the uncertainty associated with our model and relies on some arbitrary assumptions such as the fact that H_d does not vary in different gaps. Indeed, the value of $H_d \approx 4\,\mathrm{AU}$ seems to be a compromise between a slightly larger value of H_d in gaps and a smaller value in rings (see Fig. 7.6). Therefore, we interpret this result as

implying that our results do indeed show that HD 163296 is characterized by an intermediate value of $H_d \approx 2 - 6\,\mathrm{AU}$ (and take $H_d \approx 4\,\mathrm{AU}$ as our final results), but we do not explore further the exact range of values that are allowed by our χ^2 fitting.

Quite encouragingly, HD 163296 was also analyzed recently by Liu et al. (2022). The authors of that study use an analogue method to constrain the vertical structure of the disc, and try to find the best-fitting disc scale height both globally and on every disc/gap separately. In both cases, we see that the values of H_d they find are in broad agreement with the one found here. In particular, we can make a quantitative comparison with their former method, since it is essentially the same as the one used here. Transforming their parametrization of the disc thickness into values of H_d (we do this by assuming a value for the scale height of the gas component, see Sec. 7.5.1), they find that the best-fitting profile is the one with $H_d \approx 3\,\mathrm{AU}$. This value is very close to the one we find in our analysis.

7.4.2.5 MY Lup

MY Lup is a very simple disc that does not show any major substructures. The *gap-filling effect* here is thus totally absent. However, the outer edge of the disc is still subject to the same projection effect, and therefore it can be used to determine whether different scale heights produce significant differences in the intensity profile. In other terms, even the outer edge of the disc can be considered part of an "infinitely wide gap" that extends out to infinity starting from the edge of the disc.

Thanks to the high inclination of MY Lup ($i \approx 73°$), we indeed find that there is a significant difference in the predicted intensity profiles for different values of H_d. As shown in Fig. 7.6, larger H_d values correspond to profiles that are significantly shallower than the observed ones. On the other hand, small scale height ($H_d \lesssim 4\,\mathrm{AU}$) profiles present a slope that is generally compatible with data. Therefore, despite the absence of gaps, we can still use MY Lup observations to constrain its disc scale height.

7.5 Discussion

In the last section, we applied the method outlined in Sec. 7.2 to gauge the dust scale height of DSHARP discs by using the *gap-filling effect*. We have found that: (a) only $\sim 40\%$ of discs yield significant constraints on their dust scale height; (b) for the discs where these constraints are available, we find that the dust scale height (parametrized by H_d) is generally low ($H_d \lesssim 4\,\mathrm{AU}$), with almost all systems yielding only upper limits to its value. In this section, we discuss the implications of these findings, and we put our results in a broader context by comparing them with previous relevant work

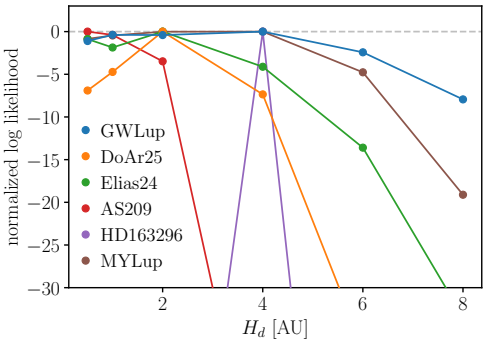

Figure 7.7: Logarithm of the likelihood ($\log \mathcal{L} \propto -\chi^2/2$) normalised to its peak value, for the values of the parameter H_d considered in our analysis. The systems discussed in Sec. 7.4 are shown with different colours.

Table 7.2: Summary of the constraints on the vertical scale-height and turbulence. Note that, as described in the main text, we have assumed $\mathrm{St} = 10^{-2}$ to break the degeneracy between α_{SS} and St.

ID	H_d [AU]	H_d/r	Θ	$\alpha_{\mathrm{SS}}/\mathrm{St}$	α_{SS}
GW Lup	$\lesssim 4$	$\lesssim 0.04$	$\lesssim 0.44$	$\lesssim 0.24$	$\lesssim 2.4 \times 10^{-3}$
DoAr 25	$\lesssim 2$	$\lesssim 0.02$	$\lesssim 0.27$	$\lesssim 0.079$	$\lesssim 7.9 \times 10^{-4}$
Elias 24	$\lesssim 2$	$\lesssim 0.02$	$\lesssim 0.20$	$\lesssim 0.042$	$\lesssim 4.2 \times 10^{-4}$
AS 209	$\lesssim 2$	$\lesssim 0.02$	$\lesssim 0.25$	$\lesssim 0.065$	$\lesssim 6.5 \times 10^{-4}$
HD 163296	≈ 4	≈ 0.04	≈ 0.56	≈ 0.45	$\approx 4.5 \times 10^{-3}$
MY Lup	$\lesssim 4$	$\lesssim 0.04$	$\lesssim 0.48$	$\lesssim 0.29$	$\lesssim 2.9 \times 10^{-3}$

on the subject. We conclude by highlighting a few caveats that need to be kept in mind when interpreting our results.

7.5.1 Relative dust and gas scale heights

As mentioned at the start of Sec. 7.2, we have focused so far only on the dust component of discs because this is the one that can be probed directly by (sub-)mm observations. However, in order to get constraints on the level of turbulence in the disc, we need to consider the vertical structure of the gas component too.

This can be done by assuming that the characteristic value of the gas scale height h_g is set by hydrostatic equilibrium ($M_\star$ is the mass of the central star):

$$h_g = \sqrt{\frac{kTr^3}{GM_\star \mu m_p}} \tag{7.6}$$

where k is the Boltzmann's constant, G is the gravitational constant, m_p is the mass of the proton, and $\mu = 2.3$ is the mean molecular weight.

Assuming the gas temperature follows the same relation we already adopted for the dust (eq. 7.3), then, we can compute the gas temperature everywhere in the disc. With the choice we have made for the radial dependence of the dust scale height (eq 7.2), it can be shown that the gas scale height follows the same dependence, and we can therefore introduce a single parameter Θ, defined as the ratio between the two scale-heights: $h_d(r) = \Theta h_g(r)$. Hereafter, we refer to the parameter Θ as the *scale height ratio*.

Given the constraints on H_d we have presented in the previous section, we can use the values of $M_\star$ and $L_\star$ given in Table 7.1 and compute the value of h_g at $r = 100\,\mathrm{AU}$, and, subsequently, the scale height ratio Θ. We list the resulting values of Θ in Table 7.2. It is easy to note how in all cases the dust scale height is smaller than the gas scale height, as expected from dust settling.

7.5.2 Implications for turbulence

The ultimate goal of this work is to put constraints on the magnitude of disc turbulence. In order to do this, we follow Dubrulle et al. (1995), who showed that

$$\Theta = \left(1 + \frac{\mathrm{St}}{\alpha_{\mathrm{SS}}}\right)^{-1/2}. \tag{7.7}$$

We list the resulting values of $\alpha_{\mathrm{SS}}/\mathrm{St}$ in Table 7.2. Note that turning these constraints into a constraint on α_{SS} requires a measurement of St, which at the moment is not available for our whole sample. In the future this may become possible through multi-wavelength observations which measure the spectral index, though significant uncertainties about the dust opacity still remain (e.g., Sierra et al. 2021; Guidi et al. 2022). For the sake of the discussion, we assume here a typical $\mathrm{St} = 10^{-2}$, but we stress this is not a measurement and this is an uncertainty that is carried over to the measurement of α_{SS}.

The first thing to note is that all our measurements are incompatible with a value of $\alpha_{\mathrm{SS}} = 10^{-2}$. This is in line with recent findings in the field that turbulence in proto-planetary discs is relatively weak (see Rosotti 2023 for a review) and also in line with theoretical expectations in the cold

conditions of proto-planetary discs, which are not capable to sustain the magneto-rotational instability (Balbus & Hawley 1991). For 3 discs, namely half of the sample where we can get constraints, we find even lower upper limits, namely that $\alpha_{\rm SS} < 10^{-3}$, reinforcing the statement that turbulence is weak in proto-planetary discs. Only for one case, HD 163296, our method provides a measurement and not only upper limits, implying that turbulence is (indirectly) detected in this disc. As already discussed, this is in line with the study of Liu et al. (2022), who found similar results.

The other aspect we can investigate with our results is whether turbulence is isotropic. In addition to HD163296, which was already discussed by Liu et al. (2022), some of our sources have also constraints on turbulence in the *radial* direction: namely AS209 (Rosotti et al. 2020), GW Lup and Elias 24 (Dullemond et al. 2018). Note that these constraints are also obtained by indirectly measuring $\alpha_{\rm SS}/{\rm St}$. Thus, a comparison between the turbulence level measured in the radial direction and in the vertical one is independent of the assumed Stokes number, St. It is notable that in all three cases the upper limit we derive on $\alpha_{\rm SS}/{\rm St}$ is *lower* than the value derived by Rosotti et al. (2020) for AS209 (0.06 with respect to 0.18 and 0.13, depending on which gap/ring we consider) or the lower limit for the range derived by Dullemond et al. (2018) for GW Lup and Elias 24 (0.3 and 0.08, respectively). At face value, this would imply that turbulence in the vertical direction is in fact weaker than in the radial direction. This could have implications regarding the debate on the origin of turbulence, since for example mechanisms like the Vertical Shear Instability (VSI, see Lesur et al. 2022 for a review) predict the opposite behaviour because they are particularly effective at lifting particles (e.g., Stoll & Kley 2016; Flock et al. 2017; Lin 2019; Dullemond et al. 2022). Note however that the opposite behaviour is found for HD163296, although the fact it is the only disc in our sample for which we are able to measure the vertical scale height may mean it is exceptionally thick. Considering the small sample size, we are not currently able to draw any conclusions on turbulence anisotropy, but this aspect should be revisited in the future with larger samples.

7.5.3 When does the method yield constraints on the scale height?

As we already discussed, for a significant fraction of our discs we are not able to get constraints on the dust scale height. It is worth asking under which conditions the method we use in this paper can give constraints. Considering the method relies on projection effects, we expect it to require discs to have moderate inclinations to be effective. On the other hand, we also expect the method to require deep gaps to work, in order to introduce an appreciable difference between models with different scale heights. On the contrary,

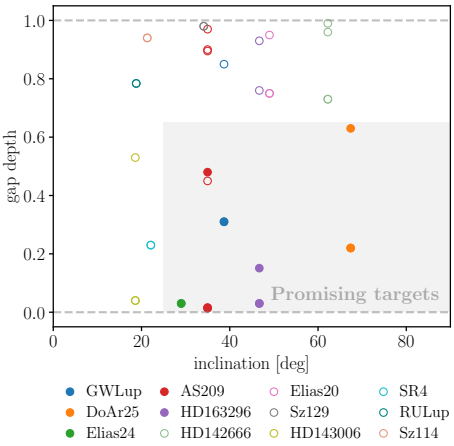

Figure 7.8: Gap depth (as defined in Huang et al. (2018)) as a function of the disc inclination angle (as reported in Tab. 7.1) for all the gaps of the DSHARP sample considered in this work for which it is possible to measure a gap depth (for more details, see Huang et al. 2018). Different colours refer to different systems in the sample. Filled (empty) circles stand for the gap that we (did not) use to measure the dust scale height effectively. The gray shaded region is defined by the two conditions inclination $> 25°$ and gap depth < 0.65, and it marks the region where we find that our method proves to be effective in constraining H_d.

shallow gaps are already filled by definition and there is less room for the *gap-filling effect* to introduce a difference between the models.

In order to quantify more our expectations, we plot in Figure 7.8 the properties of gaps in the DSHARP discs listed in Table 7.1. For every gap in these discs, we plot on the x-axis the disc's inclination, and on the y-axis the gap depth taken from Huang et al. (2018). This latter quantity is defined as the ratio between the (azimuthally-averaged) intensity in the radial bin containing the centre of the gap and the intensity in the bin containing the centre of the adjacent ring (for more details, see Huang et al. 2018). If the gap depth could not be measured, we discard the gap from our sample. We use filled (empty) circles to highlight gaps that we (did not) use to effectively constrain the dust scale height. Different systems are plotted using different colours. Note that some of the systems (i.e., AS 209, HD 163296, and GW Lup) have gaps belonging to both of these categories. This is because, in the analysis of these systems (Sec. 7.4), we have focused only on the regions where the major (i.e., deeper and larger) gaps reside. Other secondary gaps that were not considered in Sec. 7.4 are included in Fig. 7.8 with empty circles.

The figure fully confirms our expectations: gaps that can constrain the dust scale height effectively are all residing in a (gray-shaded) region for

which $i > 25°$ and the gap depth is lower than 0.65. On the other hand, gaps for which our method proves not to be effective are all outside this region, and thus they have either a small inclination or a large gap depth. The sole exception to this is a gap in AS 209 (red empty circle) which has a gap depth of ~ 0.45 and thus fall in the gray shaded region; however, this gap is located very close to the inner radius ($r \sim 9\,\mathrm{AU}$), and therefore it is likely affected by limited spatial resolution.

We stress that the criterion in which the gray shaded was defined – although it works well for our sample – is empirical and should not be taken literally. It is conceivable for example that the specific conditions may vary with the spatial resolution of the observations, as well as with the emission morphology (whose potential variation is presumably larger than what the simple gap depth parameter we introduced can catch).

Here, we have analysed only the DSHARP sample, as the largest and most homogeneous sample of high-resolution continuum observations. It is unlikely that a single programme will produce a larger sample of high-resolution observations; however, ALMA is conducting more high-resolution campaigns from many programs targeting smaller samples, and combining them one may eventually have a comparable or larger sample than the one we analysed here. The empirical criterion we have derived here may be useful for deciding which ones of those would be worth investigating using the *gap-filling effect*.

7.5.4 Caveats

The strongest caveat to make regarding this work is that we have implicitly assumed that the disc is azimuthally symmetric. The fact that strong asymmetries are relatively rare is indeed one of the main results of DSHARP (Jennings et al. 2022; Andrews et al. 2021), which partially justifies our assumption. We should caveat, however, that here we are interested in rather subtle differences in the azimuthal angle. Therefore, we cannot exclude that asymmetries are indeed present in the discs we observe, but weaker than the obvious ones such as horseshoes, crescents and spirals. This caveat is somehow mitigated by the fact that in the vast majority of cases we can only put upper limits on the dust scale-height, implying that in fact that emission is much more symmetrical (once the different spatial resolution along the major and minor axis is taken into account) than it would be if the disc were thick. The caveat remains however for the example we highlighted of DoAr 25 – where we are not able to reproduce the emission with an azimuthally symmetric disc – and for HD 163296 – where we do measure a scale-height. Though this seems unlikely, we are not able to exclude that the weak asymmetry introduced by the gap-filling effect is instead introduced by an intrinsic asymmetry, and the disc is actually thinner.

Another caveat is that we have taken here a greatly simplified disc temperature structure and we have not set up a realistic grain size distribution. This is done for the sake of simplicity; doing otherwise would introduce many other free parameters regarding the choice of dust opacity, size distribution and disc vertical structure. It is reassuring, though, that for HD163296 our method produces similar results to Liu et al. (2022), who did take the more complex route. This is probably because the method we use here is due to projection effects, and as such it should not depend directly on the details of the dust opacity or temperature.

Finally, we stress the fact that our method can gauge the value of the disc scale height (and hence of $\alpha_{\rm SS}/{\rm St}$) only locally, where substructures such as gaps and rings reside. Despite the fact that we can reproduce the observed intensity profiles everywhere, it may be that our assumption of a single, global value for the disc scale height does not correspond to reality. In principle, the vertical structure of the disc may vary from one gap to the other one; physical processes such as vortexes at the edge of gaps or meridional flows could also amplify the scale height in the proximity of gaps, biasing the inferred value of the gas turbulence. Therefore, the reader should keep in mind that our conclusions are based on a local effect, and that the knowledge of the behaviour of the scale height globally is an assumption of our model.

7.6 Summary

In this work, we have used the *gap-filling effect* to measure the dust scale height in DSHARP discs, with the goal of constraining the amount of turbulence they have. This effect originates from the fact that, in the presence of substructures such as gaps and rings, the intensity profile along the major axis differs from the profile along the minor one. This is because, if the disc inclination is not too small, line-of-sights piercing through the minor axis intercept a larger fraction of the disc's external layers – which are far from the midplane –, creating a projection effect that "fills" the gaps along that axis.

Since this effect is stronger if the disc vertical size is larger, we can probe the value of the disc scale height by building a model whose goal is to reproduce the intensity profiles along the two principal axes. Following previous work by Pinte et al. (2016) and Liu et al. (2022), we use radiative transfer to predict the resulting emission maps based on our model. The disc surface density is obtained via an iteration procedure that aims at matching the intensity observed along the major axis. This procedure is successful and convergence is reached at a satisfactory level in almost all cases (see also Fig. 7.10).

The values of the disc scale height (H_d; see eq. 7.2 for the definition) we find with our analysis can be related to the level of gas turbulence, because the vertical structure of dust grains is set by a competition between gravity and turbulence. Assuming hydrostatic equilibrium for the gas component, we can turn the value of H_d into an estimate for $\alpha_{\rm SS}/{\rm St}$, and finally into an estimate for the turbulence parameter $\alpha_{\rm SS}$ by assuming a conventional value ${\rm St} = 10^{-2}$ (see Sec. 7.5.1 for more details).

We summarise here the main findings of this paper:

- We apply our method to 15 discs from the DSHARP survey (Andrews et al. 2018). We manage to successfully constrain the value of disc scale height in 6 of these discs: GW Lup, DoAr 25, Elias 24, AS 209, HD163296, and MY Lup.

- The values of H_d we find are generally very low ($H_d \lesssim 4\,{\rm AU}$), and most estimates are upper limits only. In the single case of HD 163296, we can gauge the value of H_d to $H_d \approx 4\,{\rm AU}$ (in very good agreement with Liu et al. 2022).

- Turning these values of the disc scale height into constraints for the strength of turbulence (see Table 7.2), we find $\alpha_{\rm SS} \lesssim 5 \times 10^{-3}$. For 3 discs (i.e., half of our sample) we find even lower constraints ($\alpha_{\rm SS} < 10^{-3}$). These values are in line with recent findings that suggest a relatively low level of turbulence in protoplanetary discs (for more details, see Rosotti 2023).

- For the remaining 9 systems in our sample, we find that our method is not effective in constraining the value of the disc scale height: models with very different values of H_d give rise to identical intensity profiles along the minor axes (see Fig. 7.9). We find that all of these 9 systems ($\approx 60\%$ of our sample) are either not very inclined ($i \lesssim 25^\circ$) or they host gaps that are not deep enough – i.e., the intensity at the bottom of the gap is not much smaller than the one in the adjacent ring. We provide an empirical criterion specifying in which region of the inclination-gap depth plane (see Fig. 7.8) the method we employ here can be successfully applied.

Looking at the future, the empirical criterion we derive can be used to select from the ever-growing sample (see e.g. the catalogue assembled by Bae et al. 2022) of high-resolution disc observations those where this methodology can be applied, and in this way expand the disc sample with constraints on the vertical scale-height.

Future observations should also focus on gauging the value of the Stokes number. As we have shown in this work, the current sensitivity of observations make it possible to get good constraints on the disc vertical structure (and hence on $\alpha_{\rm SS}/{\rm St}$). However, the values we obtain for the level of gas

turbulence are subject to our lack of knowledge about the value of the Stokes number, St. Therefore, it is essential in the near future to have complementary multiwavelength observations (see e.g., Carrasco-González et al. 2019; Guidi et al. 2022) that can probe the dust grain size distribution - a sub-field that should expand in the next few years thanks to the development of band 1 on ALMA.

7.A Appendix: Discs with no constraints

We show here the results for the discs for which our method is not able to place any constraints on the value of the disc scale height. These are (see also Tab. 7.1): HD 142666, Elias 20, Sz 129, HD 143006, SR 4, RU Lup, Sz 114, WSB 52, DoAr 33. A discussion on why these systems yield no constraints on H_d is made in Sec. 7.5.3.

In Fig. 7.9, we show the intensity profiles along the minor axis extracted from data (salmon lines) together with the ones predicted by our model for different values of the disc scale height H_d (coloured lines). As it is clear from all of the plots, the reason why it is not possible to constrain H_d using our method is that all models with different values of H_d give rise to very similar profiles.

Thus, despite the fact that these profiles are generally in good agreement with data – apart from some specific cases where major asymmetries are present, e.g., the outer region of HD 143006 –, we cannot draw any conclusions on the vertical structure of the discs.

A significant exception to this is the outer region of HD 142666. Similarly to what described in the case of MY Lup (Sec. 7.4.2), profiles with a small (large) value of H_d are much (steeper) shallower due to the same projection effect that takes place in gaps and/or rings. However, in the case of HD 142666, the noise is to high to distinguish which of the different profiles is in better agreement with the data points.

7.B Appendix: Convergence along the major axis and emission maps

In this section, we show the results of our model-data comparison for what concerns the intensity profiles along the major axis (Figure 7.10) as well as the full mock images of the discs for the two extreme cases $H_d = 0.05\,\mathrm{AU}$ and $H_d = 8\,\mathrm{AU}$ (Figure 7.11- 7.12). We focus on the systems that yield significant constraints on the value of the disc scale height (see also Sec. 7.4 for more details), with the exception of GW Lup which is discussed entirely in the main text (results are in Fig. 7.5).

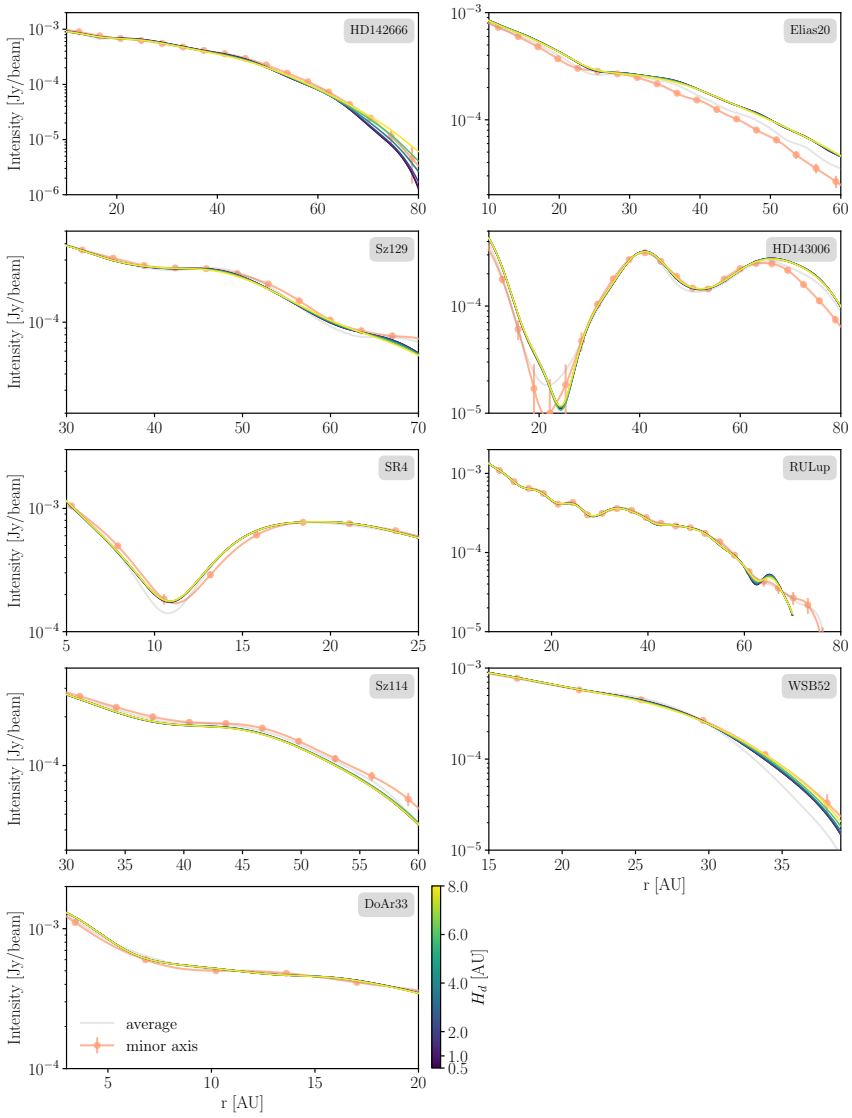

Figure 7.9: Same as Fig. 7.6, but for the systems of Tab. 7.1 for which no constraints on H_d can be placed.

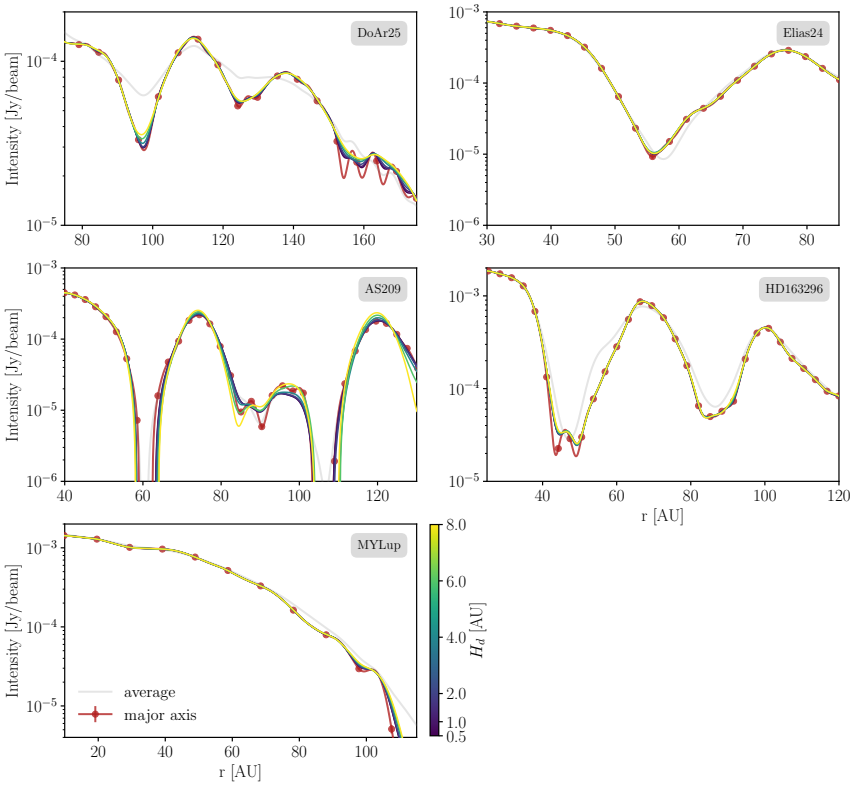

Figure 7.10: Same as the left panel of Fig. 7.5, but for the systems discussed in Sec. 7.4.2: DoAr 25 (top left), Elias 24 (top right), AS 209 (middle left), HD 163296 (middle right), MY Lup (bottom left). The same figure focusing on intensity profiles along the minor axis is in the main text (Fig. 7.6).

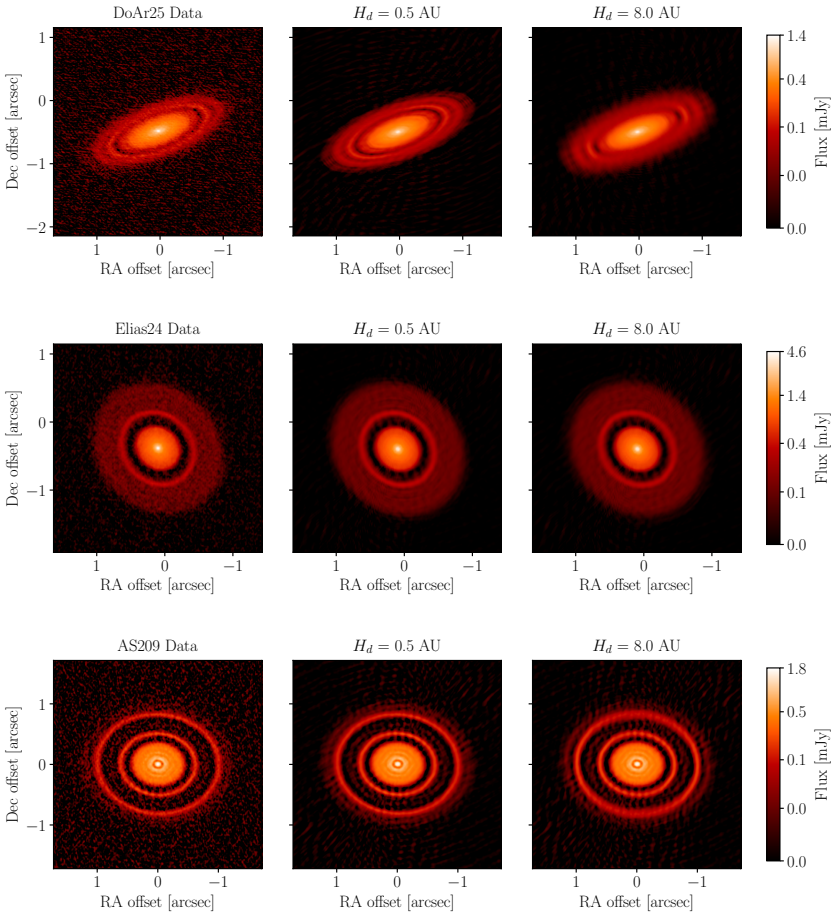

Figure 7.11: Same as Fig. 7.4, but for the systems discussed in Sec. 7.4.2 (from top to bottom): DoAr 25, Elias 24, and AS 209.

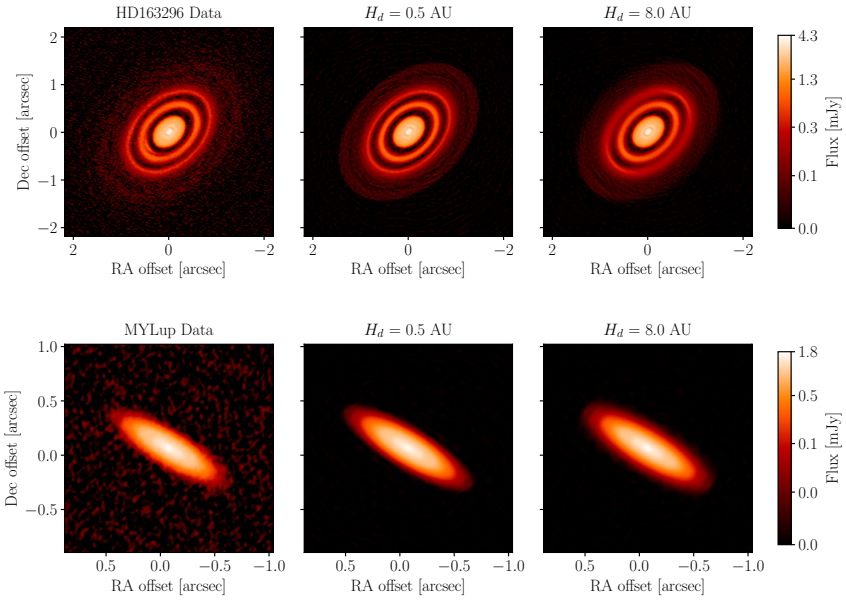

Figure 7.12: Same as Fig. 7.4 and Fig. 7.11, for the remaining systems: HD 163296 (top), MY Lup (bottom).

Acknowledgements

We thank the referee for their comments which improved the clarity of the paper. EP and GR acknowledge support from the Netherlands Organisation for Scientific Research (NWO, program number 016.Veni.192.233), as this work was originally conceived as part of the LEAPS program at Leiden Observatory. GR also acknowledges support from an STFC Ernest Rutherford Fellowship (grant number ST/T003855/1). B.T. acknowledges support from the Programme National 'Physique et Chimie du Milieu Interstellaire' (PCMI) of CNRS/INSU with INC/INP and cofunded by CNES. This work was funded by the European Union under the European Union's Horizon Europe Research & Innovation Programme grant No. 101039651 (DiscEvol). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Bibliography

Aasi J., et al., 2015, Class. Quant. Grav., 32, 074001
Abac A., et al., 2025, arXiv e-prints, p. arXiv:2503.12263
Abbott B. P., et al., 2016a, Phys. Rev. Lett., 116, 061102
Abbott B. P., et al., 2016b, Phys. Rev. Lett., 116, 131102
Abbott B. P., et al., 2018a, Living Rev. Rel., 21, 3
Abbott B. P., et al., 2018b, Phys. Rev. Lett., 120, 091101
Abbott B. P., et al., 2019a, Physical Review X, 9, 031040
Abbott B. P., et al., 2019b, Astrophys. J. Lett., 882, L24
Abbott R., et al., 2021, Astrophys. J. Lett., 913, L7
Abbott T. M. C., et al., 2022, Phys. Rev. D, 105, 023520
Abbott R., et al., 2023, Physical Review X, 13, 011048
Abramowicz M. A., Czerny B., Lasota J. P., Szuszkiewicz E., 1988, ApJ, 332, 646
Acernese F., et al., 2015, Class. Quant. Grav., 32, 024001
Adelberger K. L., Steidel C. C., 2005, ApJ, 630, 50
Agazie G., et al., 2023, ApJ, 951, L8
Aird J., et al., 2015, ApJ, 815, 66
Akins H. B., et al., 2024, arXiv e-prints, p. arXiv:2406.10341
Akiyama M., et al., 2018, PASJ, 70, S34
Akutsu T., et al., 2019, Nature Astron., 3, 35
Alexander D. M., et al., 2025, arXiv e-prints, p. arXiv:2506.19166
Algera H. S. B., et al., 2023, MNRAS, 518, 6142
Allevato V., et al., 2014, ApJ, 796, 4
Amaro-Seoane P., et al., 2023, Living Reviews in Relativity, 26, 2
Ananna T. T., Bogdán Á., Kovács O. E., Natarajan P., Hickox R. C., 2024, ApJ, 969, L18
Andika I. T., et al., 2024, A&A, 685, A25
Andrews S. M., et al., 2018, ApJ, 869, L41
Andrews S. M., et al., 2021, ApJ, 916, 51
Anglés-Alcázar D., Özel F., Davé R., Katz N., Kollmeier J. A., Oppenheimer B. D., 2015, ApJ, 800, 127
Anglés-Alcázar D., et al., 2021, ApJ, 917, 53
Angulo R. E., Hahn O., 2022, Living Reviews in Computational Astrophysics, 8, 1
Angulo R. E., Pontzen A., 2016, MNRAS, 462, L1
Antonucci R., 1993, ARA&A, 31, 473
Arita J., et al., 2023, ApJ, 954, 210
Arita J., Kashikawa N., Onoue M., Yoshioka T., Takeda Y., Hoshi H., Shimizu S., 2025, MNRAS, 536, 3677

Asgari M., Mead A. J., Heymans C., 2023, The Open Journal of Astrophysics, 6, 39
Ashton G., et al., 2019, Astrophys. J. Suppl., 241, 27
Aversa R., Lapi A., de Zotti G., Shankar F., Danese L., 2015, ApJ, 810, 74
Bañados E., et al., 2016, ApJS, 227, 11
Bañados E., et al., 2018, Nature, 553, 473
Babak S., et al., 2010, Class. Quant. Grav., 27, 084009
Bae J., Isella A., Zhu Z., Martin R., Okuzumi S., Suriano S., 2022, arXiv e-prints, p. arXiv:2210.13314
Baggen J. F. W., et al., 2024, ApJ, 977, L13
Baibhav V., Berti E., Gerosa D., Mapelli M., Giacobbo N., Bouffanais Y., Di Carlo U. N., 2019, Phys. Rev. D, 100, 064060
Baka T., Narola H., Janquart J., Samajdar A., Dietrich T., Van Den Broeck C., 2025, arXiv e-prints, p. arXiv:2507.10304
Balbus S. A., Hawley J. F., 1991, ApJ, 376, 214
Balick B., Brown R. L., 1974, ApJ, 194, 265
Ballantyne D. R., 2017a, MNRAS, 464, 613
Ballantyne D. R., 2017b, MNRAS, 464, 626
Barai P., Gallerani S., Pallottini A., Ferrara A., Marconi A., Cicone C., Maiolino R., Carniani S., 2018, MNRAS, 473, 4003
Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, ApJ, 304, 15
Barkana R., Loeb A., 2001, Phys. Rep., 349, 125
Baron D., Ménard B., 2019, Monthly Notices of the Royal Astronomical Society, 484, 5017
Barone-Nugent R. L., et al., 2014, ApJ, 793, 17
Beckwith S. V. W., Sargent A. I., Chini R. S., Guesten R., 1990, AJ, 99, 924
Begelman M. C., Dexter J., 2025, arXiv e-prints, p. arXiv:2507.09085
Begelman M. C., Blandford R. D., Rees M. J., 1980, Nature, 287, 307
Behroozi P. S., Wechsler R. H., Conroy C., 2013, ApJ, 770, 57
Behroozi P., Wechsler R. H., Hearin A. P., Conroy C., 2019, MNRAS, 488, 3143
Bekenstein J. D., 1973, ApJ, 183, 657
Bennett J. S., Sijacki D., Costa T., Laporte N., Witten C., 2024, MNRAS, 527, 1033
Benson A. J., Bower R. G., Frenk C. S., Lacey C. G., Baugh C. M., Cole S., 2003, ApJ, 599, 38
Berlind A. A., Weinberg D. H., 2002, ApJ, 575, 587
Bezanson R., et al., 2024, ApJ, 974, 92
Bhowmick A. K., DiMatteo T., Eftekharzadeh S., Myers A. D., 2019, MNRAS, 485, 2026
Bhowmick A. K., Blecha L., Torrey P., Weinberger R., Kelley L. Z., Vogelsberger M., Hernquist L., Somerville R. S., 2024, MNRAS, 529, 3768
Blecha L., Loeb A., 2008, MNRAS, 390, 1311
Bocquet S., Saro A., Dolag K., Mohr J. J., 2016, MNRAS, 456, 2361

Bodenheimer P., D'Angelo G., Lissauer J. J., Fortney J. J., Saumon D., 2013, ApJ, 770, 120
Bogdán Á., Lovisari L., Volonteri M., Dubois Y., 2018, ApJ, 852, 131
Bogdán Á., et al., 2024, Nature Astronomy, 8, 126
Bolton C. T., 1972, Nature, 235, 271
Bond J. R., Kofman L., Pogosyan D., 1996, Nature, 380, 603
Bonoli S., Marulli F., Springel V., White S. D. M., Branchini E., Moscardini L., 2009, MNRAS, 396, 423
Bonoli S., Shankar F., White S. D. M., Springel V., Wyithe J. S. B., 2010, MNRAS, 404, 399
Booth C. M., Schaye J., 2009, MNRAS, 398, 53
Booth C. M., Schaye J., 2010, MNRAS, 405, L1
Borhanian S., Sathyaprakash B. S., 2022, arXiv e-prints
Bouwens R. J., et al., 2015, ApJ, 803, 34
Bower R. G., Benson A. J., Malbon R., Helly J. C., Frenk C. S., Baugh C. M., Cole S., Lacey C. G., 2006, MNRAS, 370, 645
Bower R. G., Schaye J., Frenk C. S., Theuns T., Schaller M., Crain R. A., McAlpine S., 2017, MNRAS, 465, 32
Boyle B. J., Shanks T., Croom S. M., Smith R. J., Miller L., Loaring N., Heymans C., 2000, MNRAS, 317, 1014
Bromm V., Loeb A., 2003, The Astrophysical Journal, 596, 34
Buchner J., et al., 2015, ApJ, 802, 89
Burke-Spolaor S., et al., 2019, A&A Rev., 27, 5
CASA Team et al., 2022, PASP, 134, 114501
Caplar N., Lilly S. J., Trakhtenbrot B., 2015, ApJ, 811, 148
Carniani S., et al., 2024, Nature, 633, 318
Carrasco-González C., et al., 2019, ApJ, 883, 71
Casey C. M., et al., 2025, arXiv e-prints, p. arXiv:2505.18873
Cen R., Safarzadeh M., 2015, ApJ, 798, L38
Chandro-Gómez Á., et al., 2025, MNRAS, 539, 776
Chatziioannou K., Cornish N., Wijngaarden M., Littenberg T. B., 2021, Phys. Rev. D, 103, 044013
Chen H., et al., 2022, ApJ, 931, 29
Chiang E. I., Goldreich P., 1997, ApJ, 490, 368
Christian P., Vitale S., Loeb A., 2018, Phys. Rev. D, 98, 103022
Ciotti L., Haiman Z., Ostriker J. P., 2003, in Bender R., Renzini A., eds, The Mass of Galaxies at Low and High Redshift. p. 106 (arXiv:astro-ph/0112131), doi:10.1007/10899892_25
Circosta C., et al., 2019, A&A, 623, A172
Cole S., Kaiser N., 1989, MNRAS, 237, 1127
Cole J. W., et al., 2023, arXiv e-prints, p. arXiv:2312.10152
Comparat J., Prada F., Yepes G., Klypin A., 2017, MNRAS, 469, 4157
Conroy C., White M., 2013, ApJ, 762, 70
Cooray A., Sheth R., 2002, Phys. Rep., 372, 1

Córdova Rosado R., et al., 2024, arXiv e-prints, p. arXiv:2409.08314
Cornish N. J., 2010, arXiv e-prints
Cornish N. J., Crowder J., 2005, Phys. Rev. D, 72, 043005
Cornish N. J., Littenberg T. B., 2015, Class. Quant. Grav., 32, 135012
Cornish N. J., Porter E. K., 2007, Class. Quant. Grav., 24, 5729
Cornish N. J., Littenberg T. B., Bécsy B., Chatziioannou K., Clark J. A., Ghonge S., Millhouse M., 2021, Phys. Rev. D, 103, 044006
Correa C. A., Schaye J., Wyithe J. S. B., Duffy A. R., Theuns T., Crain R. A., Bower R. G., 2018, MNRAS, 473, 538
Costa T., 2024, MNRAS, 531, 930
Costa T., Sijacki D., Trenti M., Haehnelt M. G., 2014, MNRAS, 439, 2146
Crain R. A., van de Voort F., 2023, ARA&A, 61, 473
Croom S. M., Smith R. J., Boyle B. J., Shanks T., Miller L., Outram P. J., Loaring N. S., 2004, MNRAS, 349, 1397
Croom S. M., et al., 2005, MNRAS, 356, 415
Croton D. J., 2009, MNRAS, 394, 1109
Croton D. J., et al., 2006, MNRAS, 365, 11
Crowder J., Cornish N. J., 2004, Phys. Rev. D, 70, 082004
Crowder J., Cornish N., 2007, Phys. Rev. D, 75, 043008
Cutler C., Harms J., 2006, Phys. Rev. D, 73, 042001
D'Alessio P., Cantö J., Calvet N., Lizano S., 1998, ApJ, 500, 411
D'Amato Q., et al., 2020, A&A, 636, A37
DESI Collaboration et al., 2016, arXiv e-prints, p. arXiv:1611.00036
Dalmasso N., Trenti M., Leethochawalit N., 2024, MNRAS, 528, 898
Davies F. B., et al., 2018, ApJ, 864, 142
Davies F. B., Hennawi J. F., Eilers A.-C., 2019, ApJ, 884, L19
Davies F. B., Hennawi J. F., Eilers A.-C., 2020, MNRAS, 493, 1330
Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, Astrophysical Journal, 292, 371
Davis B. L., Graham A. W., Cameron E., 2019, Astrophysical Journal, 873, 85
Dawson K. S., et al., 2013, The Astronomical Journal, 145, 10
Dawson K. S., et al., 2016, The Astronomical Journal, 151, 44
Dayal P., Ferrara A., 2018, Phys. Rep., 780, 1
Dayal P., et al., 2024, arXiv e-prints, p. arXiv:2401.11242
DeGraf C., Sijacki D., 2017, MNRAS, 466, 3331
DeGraf C., Di Matteo T., Khandai N., Croft R., Lopez J., Springel V., 2012, MNRAS, 424, 1892
Degraf C., Di Matteo T., Springel V., 2010, MNRAS, 402, 1927
Dekel A., Birnboim Y., 2006, MNRAS, 368, 2
Devecchi B., Volonteri M., 2009, ApJ, 694, 302
Di Matteo T., Springel V., Hernquist L., 2005, Nature, 433, 604
Di Matteo T., Khandai N., DeGraf C., Feng Y., Croft R. A. C., Lopez J., Springel V., 2012, ApJ, 745, L29

Diemer B., 2018, ApJS, 239, 35
Ding X., et al., 2023, Nature, 621, 51
Dotti M., Sesana A., Decarli R., 2012, Advances in Astronomy, 2012, 940568
Drazkowska J., et al., 2022, arXiv e-prints, p. arXiv:2203.09759
Dubrulle B., Morfill G., Sterzik M., 1995, ICARUS, 114, 237
Dullemond C. P., Dominik C., Natta A., 2001, ApJ, 560, 957
Dullemond C. P., et al., 2018, ApJ, 869, L46
Dullemond C. P., Ziampras A., Ostertag D., Dominik C., 2022, A&A, 668, A105
Durodola E., Pacucci F., Hickox R. C., 2024, arXiv e-prints, p. arXiv:2406.10329
EPTA Collaboration et al., 2023, A&A, 678, A50
Eckart A., Genzel R., 1996, Nature, 383, 415
Efstathiou G., Rees M. J., 1988, MNRAS, 230, 5p
Efstathiou G., Davis M., Frenk C. S., White S. D. M., 1985, Astrophysical Journal Supplement Series, 57, 241
Eftekharzadeh S., et al., 2015, MNRAS, 453, 2779
Eilers A.-C., Davies F. B., Hennawi J. F., Prochaska J. X., Lukić Z., Mazzucchelli C., 2017, ApJ, 840, 24
Eilers A.-C., Hennawi J. F., Davies F. B., 2018, ApJ, 867, 30
Eilers A.-C., et al., 2020, ApJ, 900, 37
Eilers A.-C., Hennawi J. F., Davies F. B., Simcoe R. A., 2021, ApJ, 917, 38
Eilers A.-C., et al., 2023, ApJ, 950, 68
Eilers A.-C., et al., 2024, arXiv e-prints, p. arXiv:2403.07986
Eke V. R., et al., 2004, Monthly Notices of the Royal Astronomical Society, 355, 769
Elahi P. J., Cañas R., Poulton R. J. J., Tobar R. J., Willis J. S., Lagos C. d. P., Power C., Robotham A. S. G., 2019, PASA, 36, e021
Elbers W., Frenk C. S., Jenkins A., Li B., Pascoli S., 2021, MNRAS, 507, 2614
Endsley R., et al., 2022, MNRAS, 512, 4248
Endsley R., et al., 2023, MNRAS, 520, 4609
Endsley R., et al., 2024, MNRAS, 533, 1111
Evans M., et al., 2021, arXiv e-prints
Event Horizon Telescope Collaboration et al., 2019, ApJ, 875, L1
Event Horizon Telescope Collaboration et al., 2022, ApJ, 930, L12
Fabian A. C., 2012, ARA&A, 50, 455
Fairhurst S., Green R., Hoy C., Hannam M., Muir A., 2020, Phys. Rev. D, 102, 024055
Fan X., et al., 2006, AJ, 132, 117
Fan X., Bañados E., Simcoe R. A., 2023, ARA&A, 61, 373
Fanidakis N., Baugh C. M., Benson A. J., Bower R. G., Cole S., Done C., Frenk C. S., 2010, Monthly Notices of the Royal Astronomical Society, 410, 53

Fanidakis N., et al., 2012, MNRAS, 419, 2797
Fanidakis N., Macciò A. V., Baugh C. M., Lacey C. G., Frenk C. S., 2013, MNRAS, 436, 315
Farina E. P., et al., 2022, ApJ, 941, 106
Faucher-Giguère C.-A., 2018, MNRAS, 473, 3717
Feng Y., Di-Matteo T., Croft R. A., Bird S., Battaglia N., Wilkins S., 2016, MNRAS, 455, 2778
Ferrarese L., Merritt D., 2000, Astrophysical Journal Letters, 539, L9
Finkelstein S. L., et al., 2024, ApJ, 969, L2
Finstad D., Brown D. A., 2020, Astrophys. J. Lett., 905, L9
Flaherty K., et al., 2020, ApJ, 895, 109
Flock M., Nelson R. P., Turner N. J., Bertrang G. H. M., Carrasco-González C., Henning T., Lyra W., Teague R., 2017, ApJ, 850, 131
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, Publications of the Astronomical Society of the Pacific, 125, 306
Forouhar Moreno V. J., Helly J., McGibbon R., Schaye J., Schaller M., Han J., Kugel R., 2025, arXiv e-prints, p. arXiv:2502.06932
Fossati M., et al., 2021, MNRAS, 503, 3044
Friedmann A., 1922, Zeitschrift für Physik, 10, 377
Friedmann A., 1924, Zeitschrift für Physik, 21, 326
Furlanetto S. R., Kamionkowski M., 2006, MNRAS, 366, 529
Furtak L. J., et al., 2024, Nature, 628, 57
García-Vergara C., Hennawi J. F., Barrientos L. F., Rix H.-W., 2017, ApJ, 848, 7
García-Vergara C., Hennawi J. F., Barrientos L. F., Arrigoni Battaia F., 2019, ApJ, 886, 79
Gebhardt K., et al., 2000, Astrophysical Journal Letters, 539, L13
Gehrels N., 1986, ApJ, 303, 336
Genina A., Springel V., Rantala A., 2024, MNRAS, 534, 957
Ghez A. M., et al., 2008, The Astrophysical Journal, 689, 1044
Giallongo E., et al., 2019, ApJ, 884, 19
Gilli R., et al., 2022, A&A, 666, A17
Glikman E., Djorgovski S. G., Stern D., Dey A., Jannuzi B. T., Lee K.-S., 2011, ApJ, 728, L26
Glikman E., et al., 2018, ApJ, 861, 37
Goodman J., Weare J., 2010, Communications in Applied Mathematics and Computational Science, 5, 65
Grazian A., et al., 2023, ApJ, 955, 60
Greene J. E., Strader J., Ho L. C., 2020, Annual Review of Astronomy and Astrophysics, 58, 257
Greene J. E., et al., 2024, ApJ, 964, 39
Greengard L., Rokhlin V., 1987, Journal of Computational Physics, 73, 325
Guidi G., et al., 2022, A&A, 664, A137
Häberle M., et al., 2024, Nature, 631, 285

Habouzit M., Volonteri M., Somerville R. S., Dubois Y., Peirani S., Pichon C., Devriendt J., 2019, MNRAS, 489, 1206
Habouzit M., et al., 2021, MNRAS, 503, 1940
Habouzit M., et al., 2022, MNRAS, 509, 3015
Hahn O., Michaux M., Rampf C., Uhlemann C., Angulo R. E., 2020, MUSIC2-monofonIC: 3LPT initial condition generator, Astrophysics Source Code Library, record ascl:2008.024 (ascl:2008.024)
Haiman Z., Hui L., 2001, ApJ, 547, 27
Haiman Z., Loeb A., 2001, ApJ, 552, 459
Han J., Jing Y. P., Wang H., Wang W., 2012, MNRAS, 427, 2437
Han J., Cole S., Frenk C. S., Benitez-Llambay A., Helly J., 2018, MNRAS, 474, 604
Harikane Y., et al., 2023, ApJ, 959, 39
He W., et al., 2018, PASJ, 70, S33
Heckman T. M., Best P. N., 2014, ARA&A, 52, 589
Heger A., Fryer C. L., Woosley S. E., Langer N., Hartmann D. H., 2003, The Astrophysical Journal, 591, 288
Hellings R. W., Downs G. S., 1983, ApJ, 265, L39
Hennawi J. F., et al., 2010, ApJ, 719, 1672
Herrmann F., Hinder I., Shoemaker D., Laguna P., Matzner R. A., 2007, ApJ, 661, 430
Hickox R. C., et al., 2011, ApJ, 731, 117
Higson E., Handley W., Hobson M., Lasenby A., 2018, Statistics and Computing, 29, 891
Hopkins P. F., Hernquist L., Martini P., Cox T. J., Robertson B., Di Matteo T., Springel V., 2005, The Astrophysical Journal, 625, L71
Hopkins P. F., Hernquist L., Cox T. J., Di Matteo T., Robertson B., Springel V., 2006, ApJS, 163, 1
Hopkins P. F., Richards G. T., Hernquist L., 2007a, ApJ, 654, 731
Hopkins P. F., Lidz A., Hernquist L., Coil A. L., Myers A. D., Cox T. J., Spergel D. N., 2007b, ApJ, 662, 110
Hopkins P. F., et al., 2025, The Open Journal of Astrophysics, 8, 48
Huang J., et al., 2018, ApJ, 869, L42
Hubble E., 1925, Popular Astronomy, 33, 252
Hubble E. P., 1929, Proceedings of the National Academy of Sciences, 15, 168
Huško F., Lacey C. G., Schaye J., Nobels F. S. J., Schaller M., 2024, MNRAS, 527, 5988
Iani E., et al., 2024, arXiv e-prints, p. arXiv:2406.18207
Ikeda H., et al., 2015, ApJ, 809, 138
Inayoshi K., Ichikawa K., 2024, ApJ, 973, L49
Inayoshi K., Maiolino R., 2025, ApJ, 980, L27
Inayoshi K., Haiman Z., Ostriker J. P., 2016, MNRAS, 459, 3738
Inayoshi K., Visbal E., Haiman Z., 2020, ARA&A, 58, 27

Inayoshi K., Onoue M., Sugahara Y., Inoue A. K., Ho L. C., 2022, ApJ, 931, L25
Jeffreys H., 1946, Proceedings of the Royal Society of London Series A, 186, 453
Jenkins A., Frenk C. S., White S. D. M., Colberg J. M., Cole S., Evrard A. E., Couchman H. M. P., Yoshida N., 2001, MNRAS, 321, 372
Jennings J., Booth R. A., Tazzari M., Rosotti G. P., Clarke C. J., 2020, MNRAS, 495, 3209
Jennings J., Booth R. A., Tazzari M., Clarke C. J., Rosotti G. P., 2022, MNRAS, 509, 2780
Jeon J., Liu B., Taylor A. J., Kokorev V., Chisholm J., Kocevski D. D., Finkelstein S. L., Bromm V., 2025, ApJ, 988, 110
Jiang Y.-F., Stone J. M., Davis S. W., 2014, ApJ, 796, 106
Jing Y. P., 1998, ApJ, 503, L9
Johansen A., Lambrechts M., 2017, Annual Review of Earth and Planetary Sciences, 45, 359
Juodžbalis I., et al., 2024, MNRAS, 535, 853
Juodžbalis I., et al., 2025, arXiv e-prints, p. arXiv:2504.03551
Kaiser N., 1984, ApJ, 284, L9
Kashikawa N., et al., 2015, ApJ, 798, 28
Kashino D., Lilly S. J., Matthee J., Eilers A.-C., Mackenzie R., Bordoloi R., Simcoe R. A., 2023, ApJ, 950, 66
Kauffmann G., Haehnelt M. G., 2002, MNRAS, 332, 529
Kelley L. Z., Blecha L., Hernquist L., 2017, MNRAS, 464, 3131
Khaire V., Srianand R., 2015, MNRAS, 451, L30
Khandai N., Di Matteo T., Croft R., Wilkins S., Feng Y., Tucker E., DeGraf C., Liu M.-S., 2015, MNRAS, 450, 1349
Khrykin I. S., Hennawi J. F., McQuinn M., Worseck G., 2016, ApJ, 824, 133
Khrykin I. S., Hennawi J. F., Worseck G., 2019, MNRAS, 484, 3897
Khrykin I. S., Hennawi J. F., Worseck G., Davies F. B., 2021, MNRAS, 505, 649
Killi M., et al., 2024, A&A, 691, A52
Kim S., et al., 2009, ApJ, 695, 809
Kocevski D. D., et al., 2023, ApJ, 954, L4
Kocevski D. D., et al., 2024, arXiv e-prints, p. arXiv:2404.03576
Kokorev V., et al., 2023, ApJ, 957, L7
Kokorev V., et al., 2024a, ApJ, 968, 38
Kokorev V., et al., 2024b, ApJ, 975, 178
Kokubo M., Harikane Y., 2024, arXiv e-prints, p. arXiv:2407.04777
Kollmeier J. A., et al., 2006, ApJ, 648, 128
Kormendy J., 1988, ApJ, 335, 40
Kormendy J., Ho L. C., 2013, ARA&A, 51, 511
Koudmani S., Somerville R. S., Sijacki D., Bourne M. A., Jiang Y.-F., Profit K., 2024, MNRAS, 532, 60

Krijt S., Bosman A. D., Zhang K., Schwarz K. R., Ciesla F. J., Bergin E. A., 2020, ApJ, 899, 134
Kugel R., et al., 2023, MNRAS, 526, 6103
Kulkarni G., Worseck G., Hennawi J. F., 2019, MNRAS, 488, 1035
Labbe I., et al., 2025, ApJ, 978, 92
Lacey C., Cole S., 1993, MNRAS, 262, 627
Lacy M., Ridgway S. E., Sajina A., Petric A. O., Gates E. L., Urrutia T., Storrie-Lombardi L. J., 2015, ApJ, 802, 102
Lambrides E., et al., 2024a, arXiv e-prints, p. arXiv:2409.13047
Lambrides E., et al., 2024b, ApJ, 961, L25
Landy S. D., Szalay A. S., 1993, ApJ, 412, 64
Lapi A., Shankar F., Bosi M., Roberts D., Fu H., Varadarajan K. M., Boco L., 2025, arXiv e-prints, p. arXiv:2507.15436
Larson R. L., et al., 2023, ApJ, 953, L29
Latif M. A., Ferrara A., 2016, PASA, 33, e051
Lemaître G., 1931, MNRAS, 91, 483
Lesur G., et al., 2022, arXiv e-prints, p. arXiv:2203.09821
Li W., Inayoshi K., Qiu Y., 2021, ApJ, 917, 60
Li J., et al., 2022, ApJ, 931, L11
Li W., et al., 2024, ApJ, 969, 69
Li Z., Inayoshi K., Chen K., Ichikawa K., Ho L. C., 2025a, ApJ, 980, 36
Li J., et al., 2025b, ApJ, 981, 19
Lidz A., Hopkins P. F., Cox T. J., Hernquist L., Robertson B., 2006, ApJ, 641, 41
Lin M.-K., 2019, MNRAS, 485, 5221
Lin X., et al., 2024, arXiv e-prints, p. arXiv:2407.17570
Lin X., et al., 2025, arXiv e-prints, p. arXiv:2505.02896
Liting X., Weinstein A., Li T., Sachdev S., et al., 2014, arXiv e-prints
Littenberg T. B., 2011, Phys. Rev. D, 84, 063009
Littenberg T., Cornish N., Lackeos K., Robson T., 2020, Phys. Rev. D, 101, 123021
Liu Y., et al., 2022, arXiv e-prints, p. arXiv:2208.09230
Liu H., Jiang Y.-F., Quataert E., Greene J. E., Ma Y., 2025, arXiv e-prints, p. arXiv:2507.07190
Looser T. J., et al., 2023, arXiv e-prints, p. arXiv:2306.02470
Lupi A., Haardt F., Dotti M., Fiacconi D., Mayer L., Madau P., 2016, MNRAS, 456, 2993
Lupi A., Haiman Z., Volonteri M., 2021, MNRAS, 503, 5046
Lupi A., Trinca A., Volonteri M., Dotti M., Mazzucchelli C., 2024, A&A, 689, A128
Lusso E., Worseck G., Hennawi J. F., Prochaska J. X., Vignali C., Stern J., O'Meara J. M., 2015, MNRAS, 449, 4204
Lynden-Bell D., 1969, Nature, 223, 690
Lynden-Bell D., Rees M. J., 1971, MNRAS, 152, 461

Madau P., Haardt F., Dotti M., 2014, ApJ, 784, L38
Madau P., Giallongo E., Grazian A., Haardt F., 2024, ApJ, 971, 75
Maggiore M., et al., 2020, JCAP, 03, 050
Magorrian J., et al., 1998, Astronomical Journal, 115, 2285
Maiolino R., et al., 2024, A&A, 691, A145
Manara C. F., Ansdell M., Rosotti G. P., Hughes A. M., Armitage P. J., Lodato G., Williams J. P., 2022, arXiv e-prints, p. arXiv:2203.09930
Marasco A., Cresci G., Posti L., Fraternali F., Mannucci F., Marconi A., Belfiore F., Fall S. M., 2021, MNRAS, 507, 4274
Marshall M. A., et al., 2023, A&A, 678, A191
Martini P., 2004, in Ho L. C., ed., Coevolution of Black Holes and Galaxies. p. 169 (arXiv:astro-ph/0304009), doi:10.48550/arXiv.astro-ph/0304009
Martini P., Weinberg D. H., 2001, ApJ, 547, 12
Matsuoka Y., et al., 2018, ApJ, 869, 150
Matsuoka Y., et al., 2022, ApJS, 259, 18
Matsuoka Y., et al., 2023, ApJ, 949, L42
Matthee J., Mackenzie R., Simcoe R. A., Kashino D., Lilly S. J., Bordoloi R., Eilers A.-C., 2023, ApJ, 950, 67
Matthee J., et al., 2024a, arXiv e-prints, p. arXiv:2412.02846
Matthee J., et al., 2024b, ApJ, 963, 129
Mayer L., Kazantzidis S., Madau P., Colpi M., Quinn T., Wadsley J., 2007, Science, 316, 1874
Mazzolari G., et al., 2024, arXiv e-prints, p. arXiv:2408.15615
Mazzucchelli C., Bañados E., Decarli R., Farina E. P., Venemans B. P., Walter F., Overzier R., 2017a, ApJ, 834, 83
Mazzucchelli C., et al., 2017b, ApJ, 849, 91
McBride J., Fakhouri O., Ma C.-P., 2009, MNRAS, 398, 1858
McGreer I. D., Eftekharzadeh S., Myers A. D., Fan X., 2016, AJ, 151, 61
McGreer I. D., Fan X., Jiang L., Cai Z., 2018, AJ, 155, 131
Meacher D., Cannon K., Hanna C., Regimbau T., Sathyaprakash B. S., 2016, Phys. Rev. D, 93, 024018
Mead A. J., Verde L., 2021, MNRAS, 503, 3095
Merloni A., 2004, MNRAS, 353, 1035
Merloni A., Heinz S., 2008, MNRAS, 388, 1011
Merloni A., et al., 2014, Monthly Notices of the Royal Astronomical Society, 437, 3550
Michaux M., Hahn O., Rampf C., Angulo R. E., 2021, MNRAS, 500, 663
Mignoli M., et al., 2020, A&A, 642, L1
Milosavljević M., Merritt D., 2001, ApJ, 563, 34
Min M., Dullemond C. P., Kama M., Dominik C., 2011, ICARUS, 212, 416
Mo H. J., White S. D. M., 1996, MNRAS, 282, 347
Morselli L., et al., 2014, A&A, 568, A1
Moster B. P., Naab T., White S. D. M., 2013, MNRAS, 428, 3121
Muñoz J. B., Mirocha J., Furlanetto S., Sabti N., 2023, MNRAS, 526, L47

Murray S. G., Diemer B., Chen Z., Neuhold A. G., Schnapp M. A., Peruzzi T., Blevins D., Engelman T., 2021, Astronomy and Computing, 36, 100487
Myers A. D., et al., 2006, ApJ, 638, 622
Naidu R. P., et al., 2022, ApJ, 940, L14
Naidu R. P., et al., 2025, arXiv e-prints, p. arXiv:2503.16596
Nelson D., et al., 2019, Computational Astrophysics and Cosmology, 6, 2
Ni Y., Di Matteo T., Gilli R., Croft R. A. C., Feng Y., Norman C., 2020, MNRAS, 495, 2135
Niida M., et al., 2020, ApJ, 904, 89
Nishimichi T., et al., 2021, DarkEmulator: Cosmological emulation code for halo clustering statistics, Astrophysics Source Code Library, record ascl:2103.009 (ascl:2103.009)
Novak G. S., Ostriker J. P., Ciotti L., 2011, ApJ, 737, 26
Oesch P. A., et al., 2023, MNRAS, 525, 2864
Ohsuga K., Mori M., Nakamoto T., Mineshige S., 2005, ApJ, 628, 368
Omukai K., Schneider R., Haiman Z., 2008, The Astrophysical Journal, 686, 801
Onions J., et al., 2012, MNRAS, 423, 1200
Onken C. A., Wolf C., Bian F., Fan X., Hon W. J., Raithel D., Tisserand P., Lai S., 2022, MNRAS, 511, 572
Ono Y., et al., 2023, ApJ, 951, 72
Oogi T., Enoki M., Ishiyama T., Kobayashi M. A. R., Makiya R., Nagashima M., 2016, MNRAS, 456, L30
Owen J. E., 2014, ApJ, 790, L7
Paardekooper S.-J., Dong R., Duffell P., Fung J., Masset F. S., Ogilvie G., Tanaka H., 2022, arXiv e-prints, p. arXiv:2203.09595
Pacucci F., Loeb A., 2020, ApJ, 895, 95
Pacucci F., Loeb A., 2022, MNRAS, 509, 1885
Pacucci F., Narayan R., 2024, ApJ, 976, 96
Pacucci F., Nguyen B., Carniani S., Maiolino R., Fan X., 2023, ApJ, 957, L3
Padovani P., et al., 2017, A&A Rev., 25, 2
Pallottini A., Ferrara A., 2023, A&A, 677, L4
Pan Z., Jiang L., Fan X., Wu J., Yang J., 2022, ApJ, 928, 172
Parsa S., Dunlop J. S., McLure R. J., 2018, MNRAS, 474, 2904
Peebles P. J. E., 1980, The large-scale structure of the universe
Pérez-González P. G., et al., 2024, ApJ, 968, 4
Petiteau A., Babak S., Sesana A., de Araújo M., 2013, Phys. Rev. D, 87, 064036
Petter G. C., Hickox R. C., Alexander D. M., Myers A. D., Geach J. E., Whalen K. E., Andonie C. P., 2023, ApJ, 946, 27
Pinte C., Dent W. R. F., Ménard F., Hales A., Hill T., Cortes P., de Gregorio-Monsalvo I., 2016, ApJ, 816, 25
Pinte C., Teague R., Flaherty K., Hall C., Facchini S., Casassus S., 2022, arXiv e-prints, p. arXiv:2203.09528

Pizzati E., Hennawi J. F., Schaye J., Schaller M., 2024a, MNRAS, 528, 4466
Pizzati E., et al., 2024b, MNRAS, 534, 3155
Pizzati E., Hennawi J. F., Schaye J., Eilers A.-C., Huang J., Schindler J.-T., Wang F., 2025, MNRAS, 539, 2910
Planck Collaboration et al., 2014, A&A, 571, A1
Porciani C., Norberg P., 2006, MNRAS, 371, 1824
Porciani C., Magliocchetti M., Norberg P., 2004, MNRAS, 355, 1010
Porras-Valverde A. J., Ricarte A., Natarajan P., Somerville R. S., Gabrielpillai A., Yung L. Y. A., 2025, arXiv e-prints, p. arXiv:2504.11566
Press W. H., Schechter P., 1974, ApJ, 187, 425
Pringle J. E., 1981, ARA&A, 19, 137
Punturo M., et al., 2010, Class. Quant. Grav., 27, 194002
Quadri G., Trinca A., Lupi A., Colpi M., Volonteri M., 2025, arXiv e-prints, p. arXiv:2505.05556
Reardon D. J., et al., 2023, ApJ, 951, L6
Regimbau T., Hughes S. A., 2009, Phys. Rev. D, 79, 062002
Regimbau T., et al., 2012, Phys. Rev. D, 86, 122001
Regimbau T., Evans M., Christensen N., Katsavounidis E., Sathyaprakash B., Vitale S., 2017, Phys. Rev. Lett., 118, 151105
Reines A. E., Volonteri M., 2015, ApJ, 813, 82
Reitze D., et al., 2019a, in Bulletin of the American Astronomical Society. p. 35 (arXiv:1907.04833)
Reitze D., et al., 2019b, Bull. Am. Astron. Soc., 51, 141
Relton P., Raymond V., 2021, Phys. Rev. D, 104, 084039
Ren K., Trenti M., 2021, ApJ, 923, 110
Ren K., Trenti M., Di Matteo T., 2020, ApJ, 894, 124
Richards G. T., et al., 2006, AJ, 131, 2766
Robson T., Cornish N., 2017, Class. Quant. Grav., 34, 244002
Rosotti G. P., 2023, New Astronomy Reviews, 96, 101674
Rosotti G. P., Teague R., Dullemond C., Booth R. A., Clarke C. J., 2020, MNRAS, 495, 173
Ross N. P., et al., 2009, ApJ, 697, 1634
Ross N. P., et al., 2013, ApJ, 773, 14
Runnoe J. C., Brotherton M. S., Shang Z., 2012a, MNRAS, 422, 478
Runnoe J. C., Brotherton M. S., Shang Z., 2012b, MNRAS, 426, 2677
Sacchi A., Bogdan A., 2025, arXiv e-prints, p. arXiv:2505.09669
Sachdev S., Regimbau T., Sathyaprakash B. S., 2020, Phys. Rev. D, 102, 024051
Salpeter E. E., 1964, ApJ, 140, 796
Samajdar A., Janquart J., Van Den Broeck C., Dietrich T., 2021, Phys. Rev. D, 104, 044003
Sanders D. B., Soifer B. T., Elias J. H., Madore B. F., Matthews K., Neugebauer G., Scoville N. Z., 1988, ApJ, 325, 74

Sanders D. B., Phinney E. S., Neugebauer G., Soifer B. T., Matthews K., 1989, ApJ, 347, 29
Sargent W. L. W., Young P. J., Boksenberg A., Shortridge K., Lynds C. R., 1978, The Astrophysical Journal, 221, 731
Sathyaprakash B., et al., 2012, Class. Quant. Grav., 29, 124013
Satyavolu S., Kulkarni G., Keating L. C., Haehnelt M. G., 2023, MNRAS, 521, 3108
Sazonov S. Y., Ostriker J. P., Ciotti L., Sunyaev R. A., 2005, MNRAS, 358, 168
Scannapieco E., Silk J., Bouwens R., 2005, ApJ, 635, L13
Schaller M., et al., 2024, MNRAS, 530, 2378
Schaye J., et al., 2015, Monthly Notices of the Royal Astronomical Society, 446, 521
Schaye J., et al., 2023, MNRAS, 526, 4978
Schechter P., 1976, ApJ, 203, 297
Schindler J.-T., et al., 2019, ApJ, 871, 258
Schindler J.-T., et al., 2023, ApJ, 943, 67
Schindler J.-T., et al., 2024, arXiv e-prints, p. arXiv:2411.11534
Schmidt M., 1963, Nature, 197, 1040
Schneider D. P., et al., 2010, AJ, 139, 2360
Scholtz J., et al., 2023, arXiv e-prints, p. arXiv:2311.18731
Semenov D., Wiebe D., 2011, ApJS, 196, 25
Sesana A., Volonteri M., Haardt F., 2007, MNRAS, 377, 1711
Shakura N. I., Sunyaev R. A., 1973, A&A, 24, 337
Shankar F., 2009, New Astron. Rev., 53, 57
Shankar F., Mathur S., 2007, ApJ, 660, 1051
Shankar F., Weinberg D. H., Miralda-Escudé J., 2009, ApJ, 690, 20
Shankar F., Weinberg D. H., Shen Y., 2010a, MNRAS, 406, 1959
Shankar F., Crocce M., Miralda-Escudé J., Fosalba P., Weinberg D. H., 2010b, ApJ, 718, 231
Shankar F., et al., 2020, Nature Astronomy, 4, 282
Sharma S., 2017, ARA&A, 55, 213
Shen Y., et al., 2007, AJ, 133, 2222
Shen Y., et al., 2009, ApJ, 697, 1656
Shen Y., et al., 2013, ApJ, 778, 98
Shen X., Hopkins P. F., Faucher-Giguère C.-A., Alexander D. M., Richards G. T., Ross N. P., Hickox R. C., 2020, MNRAS, 495, 3252
Shen X., Vogelsberger M., Boylan-Kolchin M., Tacchella S., Kannan R., 2023, MNRAS, 525, 3254
Sheth R. K., Tormen G., 1999, MNRAS, 308, 119
Sheth R. K., Mo H. J., Tormen G., 2001, MNRAS, 323, 1
Sierra A., et al., 2021, ApJS, 257, 14
Sijacki D., Springel V., Di Matteo T., Hernquist L., 2007, MNRAS, 380, 877

Simpson C., Mortlock D., Warren S., Cantalupo S., Hewett P., McLure R., McMahon R., Venemans B., 2014, MNRAS, 442, 3454
Singer L. P., Price L. R., 2016, Phys. Rev. D, 93, 024013
Sinha M., Garrison L. H., 2020, MNRAS, 491, 3022
Sądowski A., Narayan R., McKinney J. C., Tchekhovskoy A., 2014, MNRAS, 439, 503
Skilling J., 2006, Bayesian Analysis, 1, 833
Soltan A., 1982, MNRAS, 200, 115
Somerville R. S., Davé R., 2015, ARA&A, 53, 51
Speagle J. S., 2020, Mon. Not. Roy. Astron. Soc., 493, 3132
Spergel D. N., et al., 2007, ApJS, 170, 377
Springel V., et al., 2005, Nature, 435, 629
Stiavelli M., et al., 2005, ApJ, 622, L1
Stoll M. H. R., Kley W., 2016, A&A, 594, A57
Stone M. A., Lyu J., Rieke G. H., Alberts S., Hainline K. N., 2023, arXiv e-prints, p. arXiv:2310.18395
Sun G., Faucher-Giguère C.-A., Hayward C. C., Shen X., 2023, MNRAS, 526, 2665
Tanaka T., Haiman Z., 2009, ApJ, 696, 1798
Tanaka T. S., et al., 2024, arXiv e-prints, p. arXiv:2412.14246
Taylor A. J., et al., 2024, arXiv e-prints, p. arXiv:2409.06772
The LIGO Scientific Collaboration the Virgo Collaboration the KAGRA Collaboration 2025, arXiv e-prints, p. arXiv:2507.08219
Thorne K. S., 1974, ApJ, 191, 507
Timlin J. D., et al., 2018, ApJ, 859, 20
Tinker J., Kravtsov A. V., Klypin A., Abazajian K., Warren M., Yepes G., Gottlöber S., Holz D. E., 2008, ApJ, 688, 709
Tinker J. L., Robertson B. E., Kravtsov A. V., Klypin A., Warren M. S., Yepes G., Gottlöber S., 2010, ApJ, 724, 878
Trainor R., Steidel C. C., 2012, in American Astronomical Society Meeting Abstracts #219. p. 154.20
Trimble V., 1995, Publications of the Astronomical Society of the Pacific, 107, 1133
Trinca A., Schneider R., Maiolino R., Valiante R., Graziani L., Volonteri M., 2023, MNRAS, 519, 4753
Trinca A., et al., 2024, arXiv e-prints, p. arXiv:2412.14248
Tucci M., Volonteri M., 2017, A&A, 600, A64
Übler H., et al., 2023, arXiv e-prints, p. arXiv:2312.03589
Ueda Y., Akiyama M., Ohta K., Miyaji T., 2003, ApJ, 598, 886
Ueda Y., Akiyama M., Hasinger G., Miyaji T., Watson M. G., 2014, ApJ, 786, 104
Urry C. M., Padovani P., 1995, PASP, 107, 803
Valentini M., Gallerani S., Ferrara A., 2021, MNRAS, 507, 1
Veale M., White M., Conroy C., 2014, MNRAS, 445, 1144

Veitch J., et al., 2015, Phys. Rev. D, 91, 042003
Vestergaard M., Osmer P. S., 2009, ApJ, 699, 800
Villenave M., et al., 2022, ApJ, 930, 11
Vitale S., Evans M., 2017, Phys. Rev. D, 95, 064052
Vito F., et al., 2018, Monthly Notices of the Royal Astronomical Society, 473, 2378
Vito F., Di Mascia F., Gallerani S., Zana T., Ferrara A., Carniani S., Gilli R., 2022, MNRAS, 514, 1672
Vogelsberger M., et al., 2014, MNRAS, 444, 1518
Vogelsberger M., Marinacci F., Torrey P., Puchwein E., 2020, Nature Reviews Physics, 2, 42
Volonteri M., 2012, Science, 337, 544
Volonteri M., Rees M. J., 2006, ApJ, 650, 669
Volonteri M., Haardt F., Madau P., 2003, ApJ, 582, 559
Volonteri M., Lodato G., Natarajan P., 2008, MNRAS, 383, 1079
Volonteri M., Silk J., Dubus G., 2015, ApJ, 804, 148
Volonteri M., Dubois Y., Pichon C., Devriendt J., 2016, MNRAS, 460, 2979
Volonteri M., Habouzit M., Colpi M., 2021, Nature Reviews Physics, 3, 732
Wang F., et al., 2019, ApJ, 884, 30
Wang F., et al., 2021, ApJ, 907, L1
Wang F., et al., 2023, ApJ, 951, L4
Wang B., et al., 2024, arXiv e-prints, p. arXiv:2403.02304
Wang B. Y., Zhou Y., Chen W., Chen N., Di Matteo T., Croft R., Bird S., Ni Y., 2025, arXiv e-prints, p. arXiv:2503.24304
Warren M. S., Abazajian K., Holz D. E., Teodoro L., 2006, ApJ, 646, 881
Wechsler R. H., Tinker J. L., 2018, ARA&A, 56, 435
Wechsler R. H., Zentner A. R., Bullock J. S., Kravtsov A. V., Allgood B., 2006, ApJ, 652, 71
Weigel A. K., Schawinski K., Caplar N., Wong O. I., Treister E., Trakhtenbrot B., 2017, ApJ, 845, 134
Weinberger R., Bhowmick A., Blecha L., Bryan G., Buchner J., Hernquist L., Hlavacek-Larrondo J., Springel V., 2025, arXiv e-prints, p. arXiv:2502.13241
Wetzel A. R., Cohn J. D., White M., 2009, MNRAS, 394, 2182
White M., 2001, A&A, 367, 27
White M., Martini P., Cohn J. D., 2008, MNRAS, 390, 1179
White M., et al., 2012, MNRAS, 424, 933
Willott C. J., Percival W. J., McLure R. J., Crampton D., Hutchings J. B., Jarvis M. J., Sawicki M., Simard L., 2005, ApJ, 626, 657
Wolfe A. M., Burbidge G. R., 1970, ApJ, 161, 419
Worseck G., Prochaska J. X., Hennawi J. F., McQuinn M., 2016, ApJ, 825, 144
Worseck G., Khrykin I. S., Hennawi J. F., Prochaska J. X., Farina E. P., 2021, MNRAS, 505, 5084

Wu Q., Shen Y., 2022, ApJS, 263, 42
Wu J., et al., 2022, MNRAS, 517, 2659
Wyithe J. S. B., Loeb A., 2003, ApJ, 595, 614
Wyithe J. S. B., Loeb A., 2009, MNRAS, 395, 1607
Wyithe J. S. B., Padmanabhan T., 2006, MNRAS, 372, 1681
Xu H., et al., 2023, Research in Astronomy and Astrophysics, 23, 075024
Yang X., Mo H. J., van den Bosch F. C., 2003, MNRAS, 339, 1057
Yang J., et al., 2016, ApJ, 829, 33
Yang J., et al., 2020, ApJ, 897, L14
Yang J., et al., 2021, ApJ, 923, 262
Yang J., et al., 2023, ApJS, 269, 27
York D. G., et al., 2000, AJ, 120, 1579
Yu Q., Tremaine S., 2002, MNRAS, 335, 965
Yue M., Fan X., Yang J., Wang F., 2021, ApJ, 921, L27
Yue M., et al., 2024a, ApJ, 966, 176
Yue M., Eilers A.-C., Ananna T. T., Panagiotou C., Kara E., Miyaji T., 2024b, ApJ, 974, L26
Yung L. Y. A., Somerville R. S., Nguyen T., Behroozi P., Modi C., Gardner J. P., 2023, arXiv e-prints, p. arXiv:2309.14408
Zackay B., Dai L., Venumadhav T., 2018, arXiv e-prints
Zel'dovich Y. B., Novikov I. D., 1967, Soviet Astronomy, 10, 602
Zhang H., Behroozi P., Volonteri M., Silk J., Fan X., Aird J., Yang J., Hopkins P. F., 2023a, arXiv e-prints, p. arXiv:2305.19315
Zhang H., Behroozi P., Volonteri M., Silk J., Fan X., Hopkins P. F., Yang J., Aird J., 2023b, MNRAS, 518, 2123
Zhang H., Behroozi P., Volonteri M., Silk J., Fan X., Aird J., Yang J., Hopkins P. F., 2023c, MNRAS, 523, L69
Zhang W., Shu X., Sun L., Shen R.-F., Dou L., Jiang N., Wang T., 2025, Nature Astronomy, 9, 702
Zheng W., et al., 2006, ApJ, 640, 574
Zormpas A., Birnstiel T., Rosotti G. P., Andrews S. M., 2022, A&A, 661, A66
Zou F., Brandt W. N., Gallo E., Luo B., Ni Q., Xue Y., Yu Z., 2024, ApJ, 976, 6
da Ângela J., et al., 2008, MNRAS, 383, 565
de Beer S., et al., 2023, MNRAS, 526, 1850
de Graaff A., et al., 2025, arXiv e-prints, p. arXiv:2503.16600
Ďurovčíková D., et al., 2024, ApJ, 969, 162
Ďurovčíková D., et al., 2025, arXiv e-prints, p. arXiv:2505.00080
van den Bosch F. C., Yang X., Mo H. J., 2003, MNRAS, 340, 771

English summary

Research in astrophysics often advances by weaving together insights from multiple domains: observations and theory, small- and large-scale phenomena, and sometimes entirely different subfields. This thesis is a modest attempt to navigate that breadth. It brings together six studies carried out with collaborators, spanning topics from the large-scale clustering of quasars and the growth of supermassive black holes, to the discovery of unexpected populations of candidate active galactic nuclei (AGN) in the era of the James Webb Space Telescope (JWST), to the challenges posed by gravitational wave astronomy and the physics of planet-forming (protoplanetary) discs. Despite their diversity, these studies share a common motivation: to develop models and methods that connect theory with observation, and to use them as tools for interpreting data that often defy expectations and sometimes force us to rethink long-standing assumptions.

Quasar clustering and supermassive black hole growth

Quasars are the visible signposts of supermassive black hole growth. Powered by gas spiraling onto supermassive black holes at the centers of galaxies, they release extraordinary amounts of energy, making them detectable across billions of light-years. Each quasar we observe reveals a black hole caught in the act of accreting matter, offering a direct view of how these giants formed and evolved when the Universe was still young.

The distribution of quasars across space is far from uniform. Some regions are densely populated with quasars, while others are empty. This *clustering* reflects the underlying pattern of dark matter halos and encodes information about the environments that fuel black hole growth. By studying quasar abundance, clustering, and luminosity together, we can begin to answer fundamental questions: How often are supermassive black holes active as quasars? How rapidly do they grow into billion-solar-mass giants? And how do quasar populations at different cosmic epochs connect into a coherent picture of black hole evolution?

The first set of studies in this thesis (*Chapters 2, 3, and 5*) tackles these questions by developing models that link quasar luminosity functions and clustering statistics, and by embedding black hole growth into the broader framework of cosmological structure formation.

Chapter 2 addresses a long-standing puzzle: quasars observed about 1.5 billion years after the Big Bang appear to cluster far more strongly than theory predicts. Using large cosmological simulations, I develop a framework that ties quasar abundance and spatial distribution to halo demographics. The results suggest that, if observations are correct, nearly all quasars at that epoch were confined to the most massive halos, with little scatter

between halo mass and quasar brightness. This implies an unusually tight link between supermassive black holes and their host dark matter halos – tighter than seen at later times – raising the question of whether quasars evolve in fundamentally different ways across cosmic history.

Chapter 3 extends this framework to even earlier epochs, when the Universe was less than a billion years old, and incorporates galaxies detected by JWST alongside quasars. Employing one of the largest cosmological simulations ever run, I reproduce key observations of both quasars and galaxies. The resulting model shows that quasars in these early epochs were active for only a tiny fraction of cosmic time. This contrasts sharply with slightly later periods, when quasars seem to shine in nearly every massive halo, and it intensifies the tension with the rapid, near-continuous growth needed to build billion-solar-mass black holes in the young Universe. Reconciling this tension remains a key challenge: are black holes truly so intermittent, and if so, how can such sporadic activity produce the enormous masses we observe?

Chapter 5 takes a first step to directly address these questions by introducing an evolutionary model that follows supermassive black holes across cosmic history. Here, black hole growth is tied directly to the assembly of their host halos, consistently tracking accretion and mergers. Despite its simplicity, the model reproduces observations from the epoch of reionization through cosmic noon, highlighting the importance of rapid, well-timed accretion episodes in producing the earliest supermassive black holes. It also provides a flexible basis for incorporating new constraints in the future.

Taken together, these studies use quasars as powerful tracers of supermassive black hole accretion and its connection to large-scale structure. They reveal both progress and open questions: quasars offer a direct view of early black hole growth, yet their activity patterns and clustering continue to challenge our understanding of how black holes evolved in the Universe.

New black hole populations in the JWST era?

The launch of JWST has opened a new window on the early Universe. With its ability to detect faint galaxies at great distances, it has revealed sources that were invisible to previous telescopes. Among the most striking discoveries is a population of compact, red objects, many showing broad emission lines typically associated with active black holes. Because of their appearance, these sources have been nicknamed "little red dots".

What makes these objects remarkable is not just their appearance, but their abundance. When corrected for obscuration, many of them appear to be just as luminous as traditional quasars, yet they are found in surveys that cover much smaller areas of the sky. This implies they must be far more common than quasars of similar brightness – an observation that, if

confirmed, would overturn the prevailing view of how frequent black holes were in the early Universe.

In my work, I compared these new sources to the well-studied population of ultraviolet-bright quasars. The results show that little red dots outnumber quasars by a large and rapidly increasing factor with redshift. Their clustering also points to a clear difference: whereas quasars occupy very massive halos and show strong spatial correlations, little red dots cluster like ordinary star-forming galaxies. This strongly suggests that little red dots are not simply obscured quasars but instead mark a distinct phase of black hole growth – or, in some cases, may not be powered by black holes at all.

Beyond quasars: overlapping gravitational wave signals and the physics of protoplanetary discs

The final two studies in this thesis step outside the realm of quasars. *Chapter 6* focuses on gravitational waves. Future detectors, such as the planned Cosmic Explorer and Einstein Telescope, will be able to detect merging black holes and neutron stars with extraordinary sensitivity. This will open new opportunities for studying the Universe, but also new challenges. One of these is that so many events will be detected that their signals will sometimes overlap in time. When this happens, standard data-analysis methods may not work properly, leading to biases in the inferred properties of the sources.

Through detailed simulations, I explored how overlapping signals affect our ability to measure the masses and other properties of merging binaries. The results show that if the mergers occur within less than about half a second of each other, the overlap can significantly bias the results. But if they are separated by more than a second, and if information from detection pipelines is used wisely, the signals can still be disentangled reliably. These findings underscore the need for new analysis strategies that can handle the complex data that third-generation detectors will provide.

Chapter 7 returns to a more familiar astrophysical setting: discs of gas and dust around young stars, the birthplaces of planets. A key question in this field is the role of turbulence, which affects everything from how gas accretes onto the star to how dust grains stick together to form planets. Measuring turbulence directly is difficult, but one promising method is to infer it from the vertical thickness of the dust layer in discs.

Using high-resolution observations from the Atacama Large Millimeter/submillimeter Array (ALMA), I developed models to infer dust thickness by comparing how gaps in the discs appear from different viewing angles. Where constraints are possible, the results indicate that the dust is confined to thin layers, pointing to low levels of turbulence and raising interesting questions about the role of turbulence in the long-term evolution of protoplanetary discs.

Nederlandse samenvatting

Onderzoek in de astrofysica boekt vaak vooruitgang door inzichten uit meerdere domeinen met elkaar te verweven: observaties en theorie, kleinschalige en grootschalige fenomenen, en soms zelfs geheel verschillende deelgebieden. Dit proefschrift is een bescheiden poging om met die breedte om te gaan. Het bundelt zes studies, uitgevoerd in samenwerking met collega's, die onderwerpen bestrijken variërend van de grootschalige clustering van quasars en de groei van superzware zwarte gaten, via de ontdekking van onverwachte populaties kandidaat-actieve galactische kernen (AGN) in het tijdperk van de James Webb-ruimtetelescoop (JWST), tot de uitdagingen van de zwaartekrachtsgolfastronomie en de fysica van planeetvormende (protoplanetaire) schijven. Ondanks hun verscheidenheid delen deze studies een gemeenschappelijke drijfveer: modellen en methoden ontwikkelen die theorie en observatie verbinden, en die inzetten als instrumenten om gegevens te interpreteren die vaak de verwachtingen tarten en ons soms dwingen langgekoesterde aannames te herzien.

De clustering van quasars en de groei van superzware zwarte gaten

Quasars zijn de zichtbare wegwijzers van de groei van superzware zwarte gaten. Aangedreven door gas dat in een spiraal naar de centra van sterrenstelsels valt, waar deze zwarte gaten zich bevinden, stoten zij enorme hoeveelheden energie uit, waardoor ze over miljarden lichtjaren heen waarneembaar zijn. Elke quasar die we observeren toont een zwart gat dat actief materie aan het accreteren is, en biedt ons zo een direct venster op hoe deze reuzen zich vormden en evolueerden toen het heelal nog jong was.

De verdeling van quasars in de ruimte is verre van uniform. Sommige regio's zijn dichtbevolkt, terwijl andere leeg zijn. Deze *clustering* weerspiegelt het onderliggende patroon van donkere-materiehalos en bevat informatie over de omgevingen die de groei van zwarte gaten voeden. Door de abundantie, clustering en lichtkracht van quasars gezamenlijk te bestuderen, kunnen we fundamentele vragen benaderen: Hoe vaak zijn superzware zwarte gaten actief als quasars? Hoe snel groeien zij uit tot reuzen van miljarden zonsmassa's? En hoe hangen quasar-populaties uit verschillende kosmische tijdperken samen in een coherent beeld van de evolutie van zwarte gaten?

De eerste reeks studies in dit proefschrift (*Hoofdstukken 2, 3 en 5*) behandelt deze vragen door modellen te ontwikkelen die quasar-lichtkrachtfuncties en clusteringstatistieken koppelen, en door de groei van zwarte gaten in te bedden in het bredere kader van de vorming van kosmische structuren.

Hoofdstuk 2 gaat in op een oud raadsel: quasars die ongeveer 1,5 miljard jaar na de Oerknal werden waargenomen blijken zich veel sterker te

clusteren dan de theorie voorspelt. Met behulp van grootschalige kosmologische simulaties ontwikkelde ik een raamwerk dat quasar-abundantie en ruimtelijke verdeling koppelt aan de eigenschappen van halos. De resultaten suggereren dat, indien de observaties correct zijn, vrijwel alle quasars in die periode beperkt waren tot de meest massieve halos, met weinig spreiding tussen halomassa en quasar-helderheid. Dit impliceert een ongewoon sterke koppeling tussen superzware zwarte gaten en hun gast-halos van donkere materie — sterker dan later in de kosmische geschiedenis — en roept de vraag op of quasars zich fundamenteel verschillend ontwikkelen doorheen de kosmische tijd.

Hoofdstuk 3 breidt dit raamwerk uit naar nog vroegere tijdperken, toen het heelal jonger was dan één miljard jaar, en neemt naast quasars ook door JWST gedetecteerde sterrenstelsels mee. Door gebruik te maken van een van de grootste kosmologische simulaties ooit uitgevoerd, reproduceerde ik belangrijke observaties van zowel quasars als sterrenstelsels. Het resulterende model toont dat quasars in deze vroege tijdperken slechts gedurende een fractie van de kosmische tijd actief waren. Dit staat in scherp contrast met iets latere perioden, waarin quasars in bijna elke massieve halo lijken te stralen, en versterkt de spanning met de snelle, vrijwel continue groei die nodig is om in het jonge heelal zwarte gaten van miljarden zonsmassa's te vormen. Deze spanning verzoenen blijft een belangrijke uitdaging: zijn zwarte gaten werkelijk zo intermitterend, en zo ja, hoe kan een zulke sporadische activiteit leiden tot de enorme massa's die we observeren?

Hoofdstuk 5 zet een eerste stap om deze vragen direct aan te pakken door een evolutionair model te introduceren dat superzware zwarte gaten doorheen de kosmische geschiedenis volgt. Hier wordt de groei van zwarte gaten rechtstreeks gekoppeld aan de opbouw van hun gast-halos, waarbij accretie en fusies consistent worden bijgehouden. Ondanks zijn eenvoud weet het model observaties te reproduceren van de reïonisatie-epoque tot aan het kosmische middaguur, en benadrukt het de rol van snelle, goed getimede accretie-episoden bij het vormen van de vroegste superzware zwarte gaten. Het biedt bovendien een flexibel uitgangspunt voor het opnemen van nieuwe waarnemingsconstraints in de toekomst.

Gezamenlijk gebruiken deze studies quasars als krachtige tracers van de accretie op superzware zwarte gaten en hun verbinding met grootschalige structuur. Ze laten zowel vooruitgang als open vragen zien: quasars bieden een direct venster op de vroege groei van zwarte gaten, maar hun activiteitspatronen en clustering blijven een uitdaging vormen voor ons begrip van de evolutie van zwarte gaten in het heelal.

Nieuwe zwartegatpopulaties in het JWST-tijdperk?

De lancering van JWST heeft een nieuw venster geopend op het vroege heelal. Dankzij zijn vermogen om zwakke sterrenstelsels op grote afstanden

te detecteren, heeft de telescoop bronnen onthuld die voorheen onzichtbaar waren voor eerdere telescopen. Tot de meest opvallende ontdekkingen behoort een populatie van compacte, rode objecten, waarvan vele brede emissielijnen vertonen die typisch worden geassocieerd met actieve zwarte gaten. Vanwege hun uiterlijk worden deze bronnen "little red dots" genoemd.

Wat deze objecten opmerkelijk maakt, is niet alleen hun verschijning, maar ook hun overvloed. Na correctie voor verduistering blijken veel van hen net zo lichtkrachtig te zijn als traditionele quasars, maar ze worden gevonden in onderzoeken die slechts veel kleinere delen van de hemel bestrijken. Dit impliceert dat ze veel talrijker moeten zijn dan quasars met een vergelijkbare helderheid — een observatie die, indien bevestigd, het heersende beeld van de frequentie van zwarte gaten in het vroege heelal volledig zou omgooien.

In mijn werk heb ik deze nieuwe bronnen vergeleken met de goed bestudeerde populatie van ultraviolet-heldere quasars. De resultaten tonen aan dat little red dots quasars in aantal ruimschoots overtreffen, met een factor die snel toeneemt bij hogere redshift. Hun clustering wijst bovendien op een duidelijk verschil: terwijl quasars zich bevinden in zeer massieve halos en sterke ruimtelijke correlaties vertonen, clusteren little red dots als gewone stervormende sterrenstelsels. Dit suggereert sterk dat little red dots niet simpelweg verduisterde quasars zijn, maar eerder een afzonderlijke fase in de groei van zwarte gaten markeren — of, in sommige gevallen, misschien helemaal niet door zwarte gaten worden aangedreven.

Voorbij quasars: overlappende zwaartekrachtsgolfsignalen en de fysica van protoplanetaire schijven

De laatste twee studies in dit proefschrift vallen buiten het domein van quasars. *Hoofdstuk 6* richt zich op zwaartekrachtsgolven. Toekomstige detectoren, zoals de geplande Cosmic Explorer en Einstein Telescope, zullen samensmeltende zwarte gaten en neutronensterren met buitengewone gevoeligheid kunnen detecteren. Dit opent nieuwe mogelijkheden om het heelal te bestuderen, maar brengt ook nieuwe uitdagingen met zich mee. Eén daarvan is dat er zoveel gebeurtenissen zullen worden waargenomen dat hun signalen soms in de tijd overlappen. Wanneer dat gebeurt, kunnen standaard data-analysemethoden niet goed meer werken, wat leidt tot vertekeningen in de afgeleide eigenschappen van de bronnen.

Met behulp van gedetailleerde simulaties heb ik onderzocht hoe overlappende signalen ons vermogen beïnvloeden om de massa's en andere eigenschappen van samensmeltende dubbelsterren te meten. De resultaten laten zien dat, als de fusies binnen minder dan ongeveer een halve seconde van elkaar plaatsvinden, de overlap de resultaten aanzienlijk kan vertekenen. Vinden ze daarentegen meer dan een seconde uit elkaar plaats, en wordt informatie uit detectiepijplijnen verstandig gebruikt, dan kunnen de signalen nog steeds betrouwbaar van elkaar worden gescheiden. Deze bevindingen

benadrukken de noodzaak van nieuwe analysemethoden die de complexe gegevens van detectoren van de derde generatie aankunnen.

Hoofdstuk 7 keert terug naar een meer vertrouwde astrofysische omgeving: schijven van gas en stof rond jonge sterren, de geboorteplaatsen van planeten. Een kernvraag in dit vakgebied is de rol van turbulentie, die alles beïnvloedt van de manier waarop gas op de ster accreteert tot hoe stofdeeltjes samenklonteren om planeten te vormen. Turbulentie direct meten is moeilijk, maar een veelbelovende methode is om die af te leiden uit de verticale dikte van de stoflaag in de schijven.

Met hogeresolutie-waarnemingen van de Atacama Large Millimeter/submillimeter Array (ALMA) heb ik modellen ontwikkeld om de stofdikte af te leiden door te vergelijken hoe gaten in de schijven verschijnen onder verschillende kijkhoeken. Waar beperkingen mogelijk zijn, wijzen de resultaten erop dat het stof opgesloten zit in dunne lagen, wat duidt op lage niveaus van turbulentie en belangrijke vragen oproept over de rol van turbulentie in de langetermijnevolutie van protoplanetaire schijven.

Riassunto in italiano

La ricerca in astrofisica progredisce spesso intrecciando contributi provenienti da diversi ambiti: osservazioni e teoria, fenomeni su piccola e grande scala, e talvolta persino ambiti del tutto differenti. Questo lavoro di tesi è un modesto tentativo di navigare queste diverse dimensioni. Riunisce sei studi condotti in collaborazione con colleghi, che spaziano dal clustering dei quasar su larga scala e dalla crescita dei buchi neri supermassicci, alla scoperta di nuove possibili popolazioni di nuclei galattici attivi (AGN) nell'era del James Webb Space Telescope (JWST), fino alle sfide poste dall'astronomia delle onde gravitazionali e dalla fisica dei dischi in cui si formano i pianeti. Nonostante la loro diversità, questi studi condividono un obiettivo comune: sviluppare modelli e metodi che colleghino teoria e osservazione, e utilizzarli come strumenti per interpretare dati che spesso contraddicono le aspettative e talvolta ci costringono a rivedere ipotesi di lunga data.

Il clustering dei quasar e la crescita dei buchi neri supermassicci

I quasar sono i segnali visibili della crescita dei buchi neri supermassicci. Alimentati dal gas che precipita verso i buchi neri supermassicci al centro delle galassie, rilasciano quantità straordinarie di energia che li rendono osservabili a miliardi di anni luce di distanza. Ogni quasar osservato rivela un buco nero colto nell'atto di accrescere materia, offrendo una prospettiva diretta su come questi giganti si siano formati ed evoluti quando l'Universo era ancora giovane.

La distribuzione dei quasar nello spazio non è affatto uniforme. Alcune regioni sono densamente popolate di quasar, altre quasi vuote. Questo fenomeno, chiamato *clustering*, riflette la struttura sottostante degli aloni di materia oscura e contiene informazioni sulle regioni che alimentano la crescita dei buchi neri. Studiando insieme la densità numerica, il clustering e la luminosità dei quasar, possiamo cercare di rispondere a domande fondamentali: con quale frequenza i buchi neri supermassicci sono attivi come quasar? Con quale rapidità crescono fino a raggiungere miliardi di masse solari? E come si collegano le popolazioni di quasar osservate in epoche cosmiche diverse in un quadro coerente di evoluzione dei buchi neri?

La prima serie di studi di questa tesi (*Capitoli 2, 3 e 5*) affronta questi interrogativi sviluppando modelli che collegano funzioni di luminosità e statistiche di clustering dei quasar, e legando la crescita dei buchi neri alla formazione delle strutture cosmiche nel modello cosmologico standard.

Il *Capitolo 2* affronta un enigma di lunga data: i quasar osservati circa 1,5 miliardi di anni dopo il Big Bang sembrano avere un livello di clustering più elevato di quanto atteso. Utilizzando grandi simulazioni cosmologiche,

ho sviluppato un modello che collega abbondanza e distribuzione spaziale dei quasar alle proprietà degli aloni di materia oscura. I risultati del modello suggeriscono che, se le osservazioni sono corrette, quasi tutti i quasar di quell'epoca risiedono negli aloni più massicci, con poca dispersione tra massa dell'alone e luminosità del quasar. Questo implica un legame sorprendentemente stretto tra i buchi neri supermassicci e gli aloni dove risiedono – molto più stretto di quanto osservato in epoche successive – e induce a domandarsi se i quasar evolvano in modi fondamentalmente diversi a diverse epoche cosmiche.

Il *Capitolo 3* estende questo modello a epoche ancora più remote, quando l'Universo aveva meno di un miliardo di anni, includendo insieme ai quasar anche le galassie rilevate da JWST. Utilizzando una delle più grandi simulazioni cosmologiche mai eseguite, ho riprodotto osservazioni chiave di entrambe le popolazioni, quasar e galassie. Il modello indica che i quasar a quei tempi erano attivi solo per una frazione minima del tempo totale di vita dell'Universo. Ciò contrasta nettamente con quanto osserviamo in epoche successive, quando i quasar sembrano brillare in quasi tutti gli aloni massicci, e accentua la tensione con la crescita rapida e quasi continua necessaria per formare così presto buchi neri di miliardi di masse solari. Riconciliare questa tensione resta una sfida cruciale: i buchi neri sono davvero così intermittenti nella loro attività e, se sì, come può una crescita così sporadica produrre le enormi masse dei buchi neri osservati nell'Universo giovane?

Il *Capitolo 5* compie un primo passo per affrontare direttamente queste domande, introducendo un modello evolutivo che segue i buchi neri supermassicci lungo tutta la storia cosmica. In questo modello, la crescita dei buchi neri è collegata direttamente allo sviluppo degli aloni in cui essi risiedono, integrando in modo coerente episodi di accrescimento e "merger". Nonostante la sua semplicità, il modello riproduce osservazioni che vanno dall'epoca della reionizzazione fino al mezzogiorno cosmico, e rivela l'importanza di episodi di accrescimento rapidi nella formazione dei primi buchi neri supermassicci. Inoltre, fornisce una base flessibile per integrare in futuro nuovi dati osservativi.

Nel loro insieme, questi studi utilizzano i quasar come utili strumenti per tracciare l'accrescimento dei buchi neri supermassicci e il loro legame con la struttura a larga scala dell'Universo. Evidenziano sia i progressi compiuti sia le questioni che rimangono ancora aperte: i quasar offrono una visione diretta della crescita precoce dei buchi neri, ma i loro "pattern" di attività e il loro clustering continuano a sfidare la nostra comprensione dell'evoluzione dei buchi neri nell'Universo.

Nuove popolazioni di buchi neri nell'era di JWST?

Il lancio del telescopio spaziale JWST ha aperto una nuova finestra sull'Universo giovane. Grazie alla sua capacità di rilevare galassie deboli a grandi

distanze, ha rivelato sorgenti che erano invisibili ai telescopi precedenti. Tra le scoperte più sorprendenti vi è una popolazione di oggetti compatti e rossicci, molti dei quali mostrano larghe linee di emissione che sono tipicamente associate a buchi neri attivi. Per il loro aspetto, queste sorgenti sono state soprannominate "little red dots" (letteralmente, "piccoli punti rossi").

Ciò che rende questi oggetti notevoli non è solo il loro aspetto, ma la loro abbondanza. Dopo aver applicato una correzione per l'attenuazione causata da gas e polveri, molti di questi oggetti risultano luminosi quanto i quasar tradizionali, eppure vengono trovati in campagne osservative che coprono aree del cielo molto più piccole. Questo implica che debbano essere molto più comuni dei quasar con pari luminosità – una conclusione che, se confermata, rivoluzionerebbe le teorie correnti sulla frequenza dei buchi neri nell'Universo primordiale.

Nel mio lavoro ho confrontato queste nuove sorgenti con la popolazione di quasar "classici", che sono luminosi nell'ultravioletto. I risultati mostrano che i *little red dots* superano in numero i quasar di un fattore molto grande che evolve rapidamente con il redshift. Anche il loro clustering rivela una chiara differenza: mentre i quasar si trovano in aloni molto massicci e sono quindi fortemente correlati spazialmente, i *little red dots* si distribuiscono come normali galassie "star-forming". Questo indica che i *little red dots* non sono semplici quasar oscurati, ma piuttosto rappresentano una fase distinta della crescita dei buchi neri – o, in alcuni casi, che non sono affatto buchi neri ma semplici galassie.

Oltre i quasar: segnali di onde gravitazionali sovrapposti e la fisica dei dischi protoplanetari

Gli ultimi due studi di questa tesi si collocano al di fuori del dominio dei quasar. Il *Capitolo 6* è dedicato alle onde gravitazionali. Futuri rivelatori, come il Cosmic Explorer e l'Einstein Telescope, saranno in grado di osservare fusioni di buchi neri e stelle di neutroni con una sensibilità straordinaria. Questo aprirà nuove opportunità per studiare l'Universo, ma porterà anche nuove sfide. Una di queste è che verranno rilevati così tanti eventi che i loro segnali talvolta si sovrapporranno nel tempo. Quando ciò accade, i metodi standard di analisi possono non funzionare, introducendo bias nelle proprietà che inferiamo per le sorgenti che si stanno fondendo.

Usando simulazioni dettagliate, ho esplorato come i segnali sovrapposti influenzino la nostra capacità di misurare masse e altre proprietà delle binarie in fusione. I risultati mostrano che, se le fusioni avvengono entro meno di mezzo secondo l'una dall'altra, la sovrapposizione può distorcere in modo significativo i risultati. Se invece avvengono a più di un secondo di distanza, e se le informazioni delle pipeline di rilevamento vengono utilizzate nel modo corretto, i segnali possono ancora essere separati in modo affidabile. Questi

risultati sottolineano la necessità di nuove strategie di analisi in grado di gestire la complessità dei dati che i rivelatori di terza generazione forniranno.

Il *Capitolo 7* ritorna a un contesto astrofisico più familiare: i dischi di gas e polvere attorno a giovani stelle, le culle in cui nascono i pianeti. Una questione centrale in questo campo è il ruolo della turbolenza del gas, che influenza tutto: dal modo in cui il gas accresce sulla stella al modo in cui i grani di polvere si aggregano per formare pianeti. Misurare direttamente la turbolenza è difficile, ma un metodo promettente è dedurla dallo spessore verticale dello strato di polvere nei dischi.

Utilizzando osservazioni ad alta risoluzione dell'Atacama Large Millimeter/submillimeter Array (ALMA), ho sviluppato modelli per stimare lo spessore della polvere confrontando come appaiono i gap nei dischi a diverse angolazioni di osservazione. Dove le condizioni lo permettono, i risultati indicano che la polvere è confinata in strati sottili, suggerendo bassi livelli di turbolenza e sollevando interrogativi importanti sul ruolo della turbolenza nell'evoluzione a lungo termine dei dischi protoplanetari.

Publications

Part of this thesis.

First Author

7. **E. Pizzati**, J. F. Hennawi, J. Schaye, A. C. Eilers, J. Huang, J. Schindler, F. Wang, *"Little Red Dots" cannot reside in the same dark matter halos as comparably luminous unobscured quasars*, Monthly Notices of the Royal Astronomical Society, Volume 539, Issue 4, June 2025, Pages 2910–2925.
doi.org/10.1093/mnras/staf660

6. **E. Pizzati**, J. F. Hennawi, J. Schaye, A. C. Eilers, J. Huang, J. Schindler, F. Wang, C. S. Frenk, W. Elbers, J. C. Helly, R. Mackenzie, J. Matthee, R. Bordoloi, D. Kashino, R. P. Naidu, M. Yue, *A unified model for the clustering of quasars and galaxies at* $z \approx 6$, Monthly Notices of the Royal Astronomical Society, Volume 534, Issue 4, November 2024, Pages 3155–3175.
doi.org/10.1093/mnras/stae2307

5. **E. Pizzati**, J. F. Hennawi, J. Schaye, M. Schaller, *Revisiting the extreme clustering of* $z \approx 4$ *quasars with large volume cosmological simulations*, Monthly Notices of the Royal Astronomical Society, Volume 528, Issue 3, March 2024, Pages 4466–4489.
doi.org/10.1093/mnras/stae329

4. **E. Pizzati**, G. P. Rosotti, B. Tabone, *Constraining turbulence in protoplanetary discs using the gap contrast: an application to the DSHARP sample*, Monthly Notices of the Royal Astronomical Society, Volume 524, Issue 2, September 2023, Pages 3184–3200.
doi.org/10.1093/mnras/stad2057

3. **E. Pizzati**, A. Ferrara, A. Pallottini, L. Sommovigo, M. Kohandel, S. Carniani, *[CII] Haloes in ALPINE galaxies: smoking-gun of galactic outflows?*, Monthly Notices of the Royal Astronomical Society, Volume 519, Issue 3, March 2023, Pages 4608–4621.
doi.org/10.1093/mnras/stac3816

2. **E. Pizzati**, S. Sachdev, A. Gupta, and B. S. Sathyaprakash. *Toward inference of overlapping gravitational-wave signals*, Physical Review D, vol. 105, no. 10, 2022.
doi.org/10.1103/PhysRevD.105.104016

1. **E. Pizzati**, A. Ferrara, A. Pallottini, S. Gallerani, L. Vallini, D. Decataldo, S. Fujimoto, *Outflows and extended [CII] haloes in high-redshift galaxies*, Monthly Notices of the Royal Astronomical Society, Volume 495, Issue 1, June 2020, Pages 160–172.
doi.org/10.1093/mnras/staa1163

Contributing Author

9. B. Ding, **E. Pizzati**, J. Schaye, J. F. Hennawi, W. McDonald, M. Schaller, *The luminosity function and clustering of bright quasars in the FLAMINGO cosmological simulations*, Monthly Notices of the Royal Astronomical Society, *to be submitted*.

8. J. Schindler, J. F. Hennawi, F. B. Davies, S. E. I. Bosman, F. Wang, J. Yang, A. C. Eilers, X. Fan, K. Kakiichi, **E. Pizzati**, R. Nanni, *A first look at quasar-galaxy clustering at* $z \simeq 7$, Astronomy&Astrophysics, *submitted*.
doi:10.48550/arXiv.2510.08455

7. S. Onorato, J. F. Hennawi, **E. Pizzati**, B. P. Venemans, A. C. Eilers, *Homogeneous measurements of proximity zone sizes for 59 quasars in the Epoch of Reionization*, Monthly Notices of the Royal Astronomical Society, *submitted*.
doi:10.48550/arXiv.2505.09676

6. J. Schindler, J. F. Hennawi, F. B. Davies, S. E. I. Bosman, R. Endsley, F. Wang, J. Yang, A. J. Barth, A. C. Eilers, X. Fan, K. Kakiichi, M. Maseda, **E. Pizzati**, R. Nanni, *A Broad-line, Low-luminosity Active Galactic Nucleus at z=7.3 Anchoring a Large Galaxy Overdensity*, Nature Astronomy, *accepted*.
doi.org/10.1038/s41550-025-02660-1

5. X. Lin, F. Wang, X. Fan, Z. Cai, J. B. Champagne, F. Sun, M. Volonteri, Y. Yang, J. F. Hennawi, E. Bañados, A. Barth, A. C. Eilers, E. P. Farina, W. Liu, X. Jin, H. D. Jun, A. Lupi, K. Kakiichi, C. Mazzucchelli, M. Onoue, Z. Pan, **E. Pizzati**, S. Rojas-Ruiz, J. Schindler, B. Trakhtenbrot, Y. Shen, M. Trebitsch, M. Zhuang, R. Endsley, R. A. Meyer, Z. Li, M. Li, M. Pudoka, W. L. Tee, Y. Wu, H. Zhang, *A SPectroscopic survey of biased halos In the Reionization Era (ASPIRE): Broad-line AGN at z=4-5 revealed by JWST/NIRCam WFSS*, The Astrophysical Journal, vol. 974, no. 147, 2024.
doi.org/10.3847/1538-4357/ad6565

4. A. C. Eilers, R. Mackenzie, **E. Pizzati**, J. Matthee, J. F. Hennawi, H. Zhang, R. Bordoloi, D. Kashino, S. J. Lilly, R. P. Naidu, R. A. Simcoe, M. Yue, C. S. Frenk, J. C. Helly, M. Schaller, J. Schaye, *EIGER*

VI. The Correlation Function, Host Halo Mass and Duty Cycle of Luminous Quasars at $z \approx 6$, The Astrophysical Journal, vol. 974, no. 275, 2024.
doi.org/10.3847/1538-4357/ad778b

3. L. Sommovigo, A. Ferrara, S. Carniani, A. Pallottini, P. Dayal, **E. Pizzati**, M. Ginolfi, V. Markov, A. Faisst, *A new look at the infrared properties of $z \sim 5$ galaxies*, Monthly Notices of the Royal Astronomical Society, Volume 517, Issue 4, December 2022, Pages 5930–5941.
doi.org/10.1093/mnras/stac2997

2. Y. Fudamoto, R. Smit, R. A. A. Bowler, P. A. Oesch, R. Bouwens, M. Stefanon, H. Inami, R. Endsley, V. Gonzalez, S. Schouws, D. Stark, H. S. B. Algera, M. Aravena, L. Barrufet, E. da Cunha, P. Dayal, A. Ferrara, L. Graziani, J. A. Hodge, A. P. S. Hygate, A. K. Inoue, T. Nanayakkara, A. Pallottini, **E. Pizzati**, R. Schneider, L. Sommovigo, Y. Sugahara, M. Topping, P. van der Werf, M. Bethermin, P. Cassata, M. Dessauges-Zavadsky, E. Ibar, A. L. Faisst, S. Fujimoto, M. Ginolfi, N. Hathi, G. C. Jones, F. Pozzi, and D. Schaerer, *The ALMA REBELS Survey: Average [C II] 158 μm Sizes of Star-forming Galaxies from $z \approx 7$ to $z \approx 4$*, The Astrophysical Journal, vol. 934, no. 2, 2022.
doi.org/10.3847/1538-4357/ac7a47

1. A. Pallottini, A. Ferrara, S. Gallerani, C. Behrens, M. Kohandel, S. Carniani, L. Vallini, S. Salvadori, V. Gelli, L. Sommovigo, V. D'Odorico, F. Di Mascia, **E. Pizzati**, *A survey of high-z galaxies: SERRA simulations*, Monthly Notices of the Royal Astronomical Society, Volume 513, Issue 4, July 2022, Pages 5621–5641.
doi.org/10.1093/mnras/stac1281

Curriculum Vitae

I was born in Abano Terme (Italy) on October 31st, 1997, the first of five children. I grew up in Mira, a small town in northeastern Italy along the Riviera del Brenta, the historic waterway connecting Padova to Venezia. I was the smallest kid in sight – shy, timid, and scared of just about everything. At the same time, I was full of energy and curiosity, and I managed to do well in school while being appreciated by both classmates and teachers. In high school – finally an actual person, rather than the weird creature one becomes in middle school – I spent some genuinely happy years at Liceo Scientifico Galileo Galilei in Dolo. My feelings about school remain mixed: on one hand, it was often boring and I rarely paid attention in class; on the other, I loved the thrill of learning and discovering the world. I was especially drawn to mathematics, physics, philosophy, and literature. A few professors – those I actually listened to – left a profound mark on me. Among them was my Italian literature teacher, Lucia Tosi. Though she spent little time with us, as she was battling cancer during our final years, she taught me to love the world in both its beauty and its suffering. In essence, she was a poet, and whatever there is that is poetic in me, she helped awaken.

In high school, though, it was often outside the classroom that I lived my most enriching experiences. The Scout association taught me lessons no chalkboard ever could, and entertaining children in my parish gave me both confidence and joy. Fueled by curiosity and by a tendency not to be content with what was given, I threw myself into a range of extracurricular activities. The ones I remember most fondly are the national Physics and Astronomy Olympiads, the philosophy debate tournament organized by the University of Padova, and a few small research internships at the University of Padova, the Osservatorio Astronomico di Asiago, and the Haus der Astronomie in Heidelberg. Through these experiences, my interest in physics and astronomy steadily grew.

Finally, I chose to study physics at the university and was admitted to the Scuola Normale Superiore (SNS) in Pisa. Passing the admission test at SNS will always remain one of the hardest – and most unexpected – accomplishments of my life. In 2016 I moved to Pisa and, as required, attended courses both at SNS and at the University of Pisa, where I obtained my Bachelor's and Master's degrees. SNS gives a lot and takes a lot: it gave me a solid foundation in mathematics and physics, some of the best teachers I will ever meet, and, most importantly, a group of brilliant people with whom I shared everything – from endless hours of study and problem-solving to water balloon wars, 24-hour table-soccer marathons, and much more. What it demanded, essentially, were several years of my life. Time, interests, social connections – everything was confined within the walls of the Scuola.

I often wonder whether it was worth it. Perhaps the answer will take time to emerge, but I know I am deeply grateful for the unique experience I was able to live.

During my studies, I had the opportunity to take part in several research experiences abroad, including a wonderful summer in Leiden as part of the LEAPS program. The project I began there with Giovanni Rosotti and Benoît Tabone eventually led to the publication that now forms one of the chapters of this dissertation. For my Bachelor's thesis, I decided to reach out to Andrea Ferrara, who leads the Cosmology Group at SNS – a choice that proved extremely rewarding. While Bachelor's theses in Italy are usually limited to short in-depth reviews of lecture topics, Andrea instead presented me with some data to explain and an idea for how to tackle the problem. I was hooked. That project became my first publication and later grew into my Master's thesis, which I completed in 2021 under the supervision of Andrea Ferrara and Andrea Pallottini, earning special praise from the examining committee (the so-called *abbraccio accademico*) as well as two national-level prizes.

Thanks to the projects I undertook during my university years, I became convinced that I wanted to explore the world of research further. I decided to return to Leiden to pursue my Ph.D. under the joint supervision of Joe Hennawi and Joop Schaye. Over these four years, the projects I worked on, the guidance of my supervisors and mentors, and the inspiring environment of the Observatory and the institutes I visited have transformed me from a good student into a scientist. I now feel that academia – with all its flaws and limitations – is the place where I belong at this stage of my life. After receiving my doctorate, I will move to the Center for Astrophysics at Harvard University to continue my research on quasars and the growth of supermassive black holes as an NHFP Einstein Fellow.

Acknowledgements

To my supervisors, Joe and Joop: I have learned more from you than I could ever put into words. You showed me not only what it takes to be an astronomer, but how to approach research with enthusiasm and curiosity. Working with you was demanding, yet always rewarding—and, above all, fun. It was exactly what I needed. Matthieu, you often stepped in as a third supervisor; I'm grateful for your help, your insights, and your kindness.

To all the other mentors who have shaped my path over these years. Christina, collaborating with you was one of the highlights of my Ph.D. Thank you for the care you showed during my visit to MIT, and for always offering encouragement and perspective. Andrea, I am deeply grateful for your support and advice throughout these years. Having you to look up to gave me the confidence to face each step in my career with more courage and trust. Giovanni, if working on protoplanetary discs in the middle of a Ph.D. on quasars ever felt manageable, it was thanks to your patience and guidance. And on the other side: Boyi, you were a fantastic Master's student—we both grew a lot through the journey we shared.

Being part of two groups was a gift—it broadened my perspective and made each day richer. Riccardo, JT, and Caitlin—thank you for the many ways you supported me; you all were deeply missed in the group. Timo, Daming, Ben, and Lars—it was a pleasure to share this journey with you. And Timo, getting to know you better over time—through climbing, skiing, and more—was something I greatly enjoyed. To the gang at UCSB: meeting you in person after so many online (and inevitably awkward) interactions was wonderful. Joey and Roi, I often felt you were like older brothers I could look up to academically. Victor and Rob, thank you for your patience with all my HBT-related questions. Evgenii, Filip, Jeger, Orestis, Will, Yunhao, and Zorry—I learned a great deal from each of you.

Who would have thought that leaving Italy I'd end up finding some of my favorite Italians? Ale, the way you connect with people is rare. You've changed me in many ways; I'd call it bullying, you'd call it painting on a blank canvas—but I know we'll both miss it dearly. Nicco, I admire the passion and enthusiasm you bring to everything you do—it's contagious. Nicole, your humor never fails to make me laugh—it's always more fun when you're there. Karin, your energy was a breath of fresh air this past year—thank you for bringing it into my life. Bianca and Paola, I'll thank you together so I don't mix up the names, but each of you made my time here warmer and brighter. Riccardo, Arianna, Luca, Elisa, Joan: thank you for the furniture hunts, the moves, the dinners, and everything in between.

These past four years have been the most intense of my life, filled with experiences that will stay with me forever: (mud)runs, ski trips, wipeouts,

canal swims, camping nights, football tournaments, and more. To everyone who helped turn those moments into memories, thank you. Andrew, whether trying a new sport or doing something ordinary, it was always better with you around. Josh, thank you for dragging me into so many things, and for being the one I both dreaded and loved to find in my office. A special thanks to the climbers—Amadeo, Thijs, Erik, Billy, Darìo, Manuel, Yuze—and to the members of the amazing Cosmos United team. To all the PhSki organizers: you created a truly wonderful tradition and made me fall in love with skiing again. And beyond Leiden, thank you Dominika, Teo, and Kai-Feng for including me in your adventures. I look forward to many more!

But life isn't just about big experiences—it's also built on the small, everyday moments. To my office mates, past and present—Piyush, Beth, Filip, Yannick. Piyush, a special thanks for your constant support, and for all the gossip. And finally, to everyone who made the observatory feel more than just a workplace: Alberto, Alfred, Amy, Andres, Anna, Anniek, Ani, Ben, Brigitte, Casper, Celine, Chloe, Christiaan, Christian, Ciaran, Dennis, Dilovan, Elena, Esther, Fran, Fraser, Gijs, Ivana, Jelle, Jessica, Julia, Jurjen, Kevin, Kirsty, Kostantinos, Leoni, Logan, Louis, Lucie, Luna, Mantas, Marta, Martje, Naadiyah, Osmar, Pavel, Pranjal, Richelle, Roland, Sam, Sid, Sill, Thomas, Veronica, Victorine, Willeke, Zeynab—thank you for all the little things we shared.

To the friends both near and far: Seyma and Vanesa, I'm so glad our paths have crossed again in many ways, and I know they will in the future. Paolo, Marco, Max, Edo, Fede, Beppe, Ale, Luca, and Lavi—the chats, travels, and video calls we've shared over the past four years have been like carrying a piece of Pisa with me. And Albe, Gimmy, and Veronica: you actually were that piece of Pisa in the Netherlands. Without all of you, I would have missed that part of my life so much more. Don and Jet, thank you for your constant support; visiting you again in State College was a special moment. And to all my lifelong friends, especially Marco: having people to return to is what gives a place its meaning.

Mamma and Papà, without you, I would never have had the courage to come this far: you are my hidden strength. Emma, Chiara, Marco, and Paolo: writing all your names costs me precious words, but I owe it to you—I'm deeply proud of each of you, and happy for the joy we now find in the time we share together. A cugini, zie, e zii: thank you for making our family special, I feel lucky to have you all in my life. Ai miei nonni, a Marisa e a Giampaolo, Leda, e Gino che non sono più qui: mi avete accompagnato e cresciuto con un amore incommensurabile. Questa tesi è dedicata a voi.

Silvia. You gave me what I never thought I'd find: a place I can call home, someone with whom I can be fully myself. Whatever the road ahead brings, I know we'll find the strength to walk it together—even from afar. I love you.