The Void Clustering of Lyα Emitters as a Probe of Reionization

by

Lucia Alexandra Perez

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2022 by the
Graduate Supervisory Committee:

Sangeeta Malhotra, Co-Chair
Nathaniel Butler, Co-Chair
James E. Rhoads
Christopher Groppi
Evan Scannapieco

ARIZONA STATE UNIVERSITY

May 2022

ABSTRACT

The distribution of galaxies traces the structure of underlying dark matter, and carries signatures of both the cosmology that evolved the universe as well as details of how galaxies interact with their environment and each other. There are many ways to measure the clustering of galaxies, each with unique strengths, uses, theoretical background, and connection to other physical concepts. One uncommon clustering statistic is the Void Probability Function (VPF): it simply asks, how likely is a circle/sphere of a given size to be empty in your galaxy sample? Simple and efficient to calculate, the VPF is tied to all higher order volume-averaged correlation functions as the 0th moment of count-in-cells, and encodes information from higher order clustering that the robust two-point correlation function cannot always capture. Using simulations of Lyman-alpha emitting galaxies across either redshift history or the epoch of reionization, this work asks: how powerful is the VPF itself? When can and should it be used for galaxy clustering? What unique constraints or guidelines can it give for the pacing of reionization, in the Lyman-α Galaxies in the Epoch of Reionization (LAGER) narrowband survey or across the Roman Space Telescope grism? This work provides practical guidelines for creating and carrying out clustering studies using the the VPF, and motivates the use of the VPF for reionization. The VPF of LAEs can complement LAGER constraints for the end of reionization, and thoroughly inform the timing and pace of reionization with Roman.

DEDICATION

For all those who have loved me, believed in me, and supported me

through this long journey.

I am grateful to have been encouraged and empowered in this path,

and to keep making you proud.

ACKNOWLEDGMENTS

As a student of ASU, I acknowledge that the Tempe campus sits on the ancestral homelands of those American Indian tribes that have inhabited this place for centuries, including the Akimel O'odham (Pima) and Pee Posh (Maricopa) peoples. As an astronomer, I acknowledge that many of our telescopes also sit on ancestral homelands and revered locations for indigenous peoples, in particular those of Hawai'i, the Atacama Desert, and southern Arizona.

Firstly, I wish to thank and acknowledge my PhD advisors, Sangeeta and James. This journey has been long and taken us many places, and you have always pulled through for me in exceptional, humbling, and touching ways when it most mattered. Thank you for seeing my potential and giving me the space and push to meet it. Thank you for giving me the space and support to find the path in this field that best worked for me. Finally, thank you for seeing the wholeness of my humanity, and your patience and kindness as I've learned how to be a scientist. To our lab group, as it's evolved through these years and labs, I thank you for the camaraderie, friendship, and cool scientific collaboration.

I give great thanks to my Ph.D. committee for their attention and support, through all the challenges of overseeing a student from thousands of miles away. I also wish to acknowledge and thank the administrative and business staff across SESE, as well as SESE faculty who aided my journey in their administrative leadership roles. These folks include but are not limited to: R. Arrowsmith, H. Hartnett, J. Garcia, R. Polley, A. Ramos. Exceptional thanks go to R. Dial, who in my head I refer to as "she from whom all blessings flow." Finally, I would like to thank now-Dr. Heffern for her hard work in making this LaTeX document functional and available.

I also wish to thank the scientific and computing community at NASA Goddard Space Flight Center. Many thanks and great acknowledgements go to the technology

and software specialists like D. Friendlander and D. Pereira, who single-handedly kept my 'NASA-fied' ASU-owned desktop running over several years. I am also very grateful for the many diverse astronomers that welcomed me into the unique Goddard community, including but not limited to J. Rigby, A. Heard, J. Kruk, and J. Gardner.

To my many SESE friends, thank you for the laughs over homework, research conferences, and nights burning off steam. Particular hugs to (in alphabetical order): Alana, Dan, Ed, Genevieve, J'Neil, Lena, Megan, Nivedita, and Tyler. And to new dear friends on the East Coast, that made the big moves full of joy and adventure: Carley, Sean, and Xiaoshan. To the friends that have stuck through and encouraged me throughout: Becca, Kristina, Prerana, Victoria, and Caleb. To Valentina and Victoria for being fantastic friends as well as my sister and cousin. Finally, one million thanks to Jessica for being an incomparably amazing friend this last year, and for bringing tons of laughter and Tsaya the cat into our apartment.

Finally, I send great love to my family who have supported through these years. Firstly, to my parents and younger sister for having always treated my journey and hopes as if they were spectacular things that there was no doubt I could achieve, even when the science kind of eluded them. To my grandmother and older siblings for instilling a legacy of brilliance, hard work, and always prioritizing a loving life. To my many cousins, aunts, and uncles, for celebrating my achievements and filling my life with laughter and adventure. *Me encanta nuestra familia, y los quiero tanto.*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

The central thesis of this work is the exploration and application of one uncommon method of quantifying galaxy clustering: the **Void Probability Function**. It simply asks: how likely is a given region to be empty in a sample? It is by definition tied to higher order clustering correlations (White, 1979; Maurogordato and Lachieze-Rey, 1987), and at its most popular, was used to measure the clustering of very small galaxy samples (Palunas *et al.*, 2004; Kashikawa *et al.*, 2006); as an additional probe dark matter halo occupation modeling (e.g. Fry 1986; Maurogordato and Lachieze-Rey 1987; Little and Weinberg 1994; Berlind and Weinberg 2002); and to define low-redshift galaxies' gravitational clustering moments (e.g. Croton *et al.* 2004b; Conroy *et al.* 2005). In recent years, the VPF has been slowly re-gaining popularity, particularly to study reionization.

The epoch of reionization (EoR) refers to a key transition of the universe: when the opaque 'fog' of neutral hydrogen that filled the post-recombination universe since the emission Cosmic Microwave Background was ionized by the light of very early galaxies. Figure 1.1 summarizes the current consensus of the EoR's place in cosmic history. The EoR was the last significant phase transition of the universe, and has much still not understood about it. Exactly when did it occur, and how quickly? What objects contributed and in what way? The EoR is difficult to simulate with how many uncertainties exist in our modeling methods for radiative transfer and more, and requires thoughtful observations (e.g. with the 21cm transition line of neutral hydrogen with very sensitive radio telescopes, or hunting for the rare galaxies
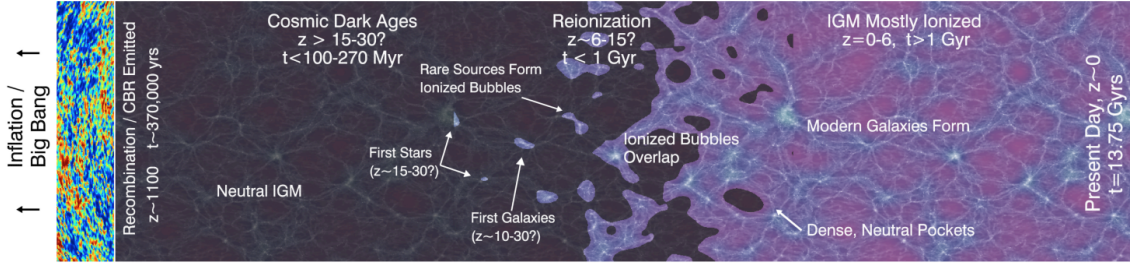
Figure 1: Cosmic Reionization The transition from the neutral intergalactic medium (IGM) left after the universe recombined at z~1100 to the fully ionized IGM observed today is termed *cosmic reionization*. After recombination, when the cosmic background radiation (CBR) currently observed in microwaves was released, hydrogen in the IGM remained neutral until the first stars and galaxies[2,4] formed at z~15-30. These primordial systems released energetic ultraviolet photons capable of ionizing local bubbles of hydrogen gas. As the abundance of these early galaxies increased, the bubbles increasingly overlapped and progressively larger volumes became ionized. This reionization process completed at z~6-8, approximately 1 Gyr after the Big Bang. At lower redshifts, the IGM remains highly ionized through radiation provided by star-forming galaxies and the gas accretion onto supermassive black holes that powers quasars.

Figure 1.1: From Robertson *et al.* (2010): an illustration of the IGM's transformation through cosmic history.

in the midst of ionizing the intergalactic medium around them).

One already fruitful probe for reionization are *Lyman-α Emitters* (LAEs). These are galaxies detected (often solely) by very bright emission in the Lyα line (the spectral line moving from $n = 2$ orbital to the $n = 1$ ground state of hydrogen, emitted during active star formation). Lyα is emitted in the UV (1216Å), meaning that as the galaxy emission becomes redshifted due to the Hubble expansion of the universe, Lyα from very early eras of the universe becomes visible in the optical (for $2 < z < 5$) or infrared ($z > 6$). See Figure 1.2 for examples of some high-redshift LAE spectra from group members. Additionally, because Lyα is the resonant line of hydrogen, it is absorbed by almost any amount of neutral hydrogen in its vicinity. Combine all these together–star forming galaxies emitting light perfectly tuned to neutral hydrogen that can be observed in the infrared–leads to the use of LAEs as probes for the EoR.

This dissertation shares one central narrative: the VPF of LAEs, with the goal of using this clustering to learn more about the EoR. In this introduction, I will
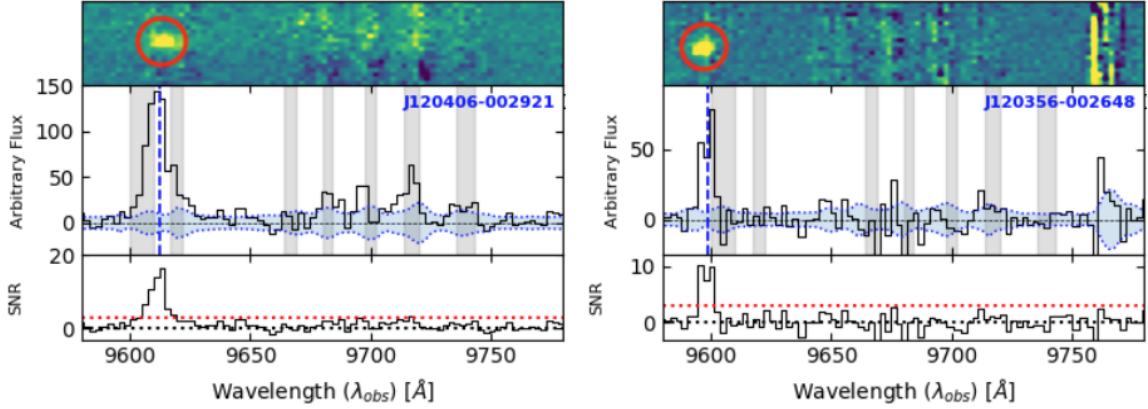
Figure 1.2: Example spectra of particularly bright LAGER $z = 6.9$ LAEs, from Harish *et al.* (2021) using the Low Resolution Imaging Spectrometer (LRIS) on the Keck I telescope in Hawai'i.

contextualize the some popular statistics used for galaxy clustering and the VPF's place among them, to help inform later chapters. Finally, I will outline the Chapters of this dissertation and what they explore about the VPF, LAEs, and constraints for reionization.

1.1 Galaxy Clustering Statistics

Dark matter is the dominant force behind the large-scale structure of galaxies and is a key component of cosmology. Galaxies trace the structure of dark matter, and their distribution in space is heavily influenced by features of dark matter, as well as signatures of how galaxies interact with their environment and each other (e.g. Peebles 1980; and Figures 1.3 and 1.4). Though we primarily focus on using the clustering of LAEs to constrain reionization in this dissertation, applications for galaxy clustering are wide-reaching.

There are many ways to measure the clustering of galaxies, each with unique strengths, uses, theoretical background, and connection to other physical concepts.
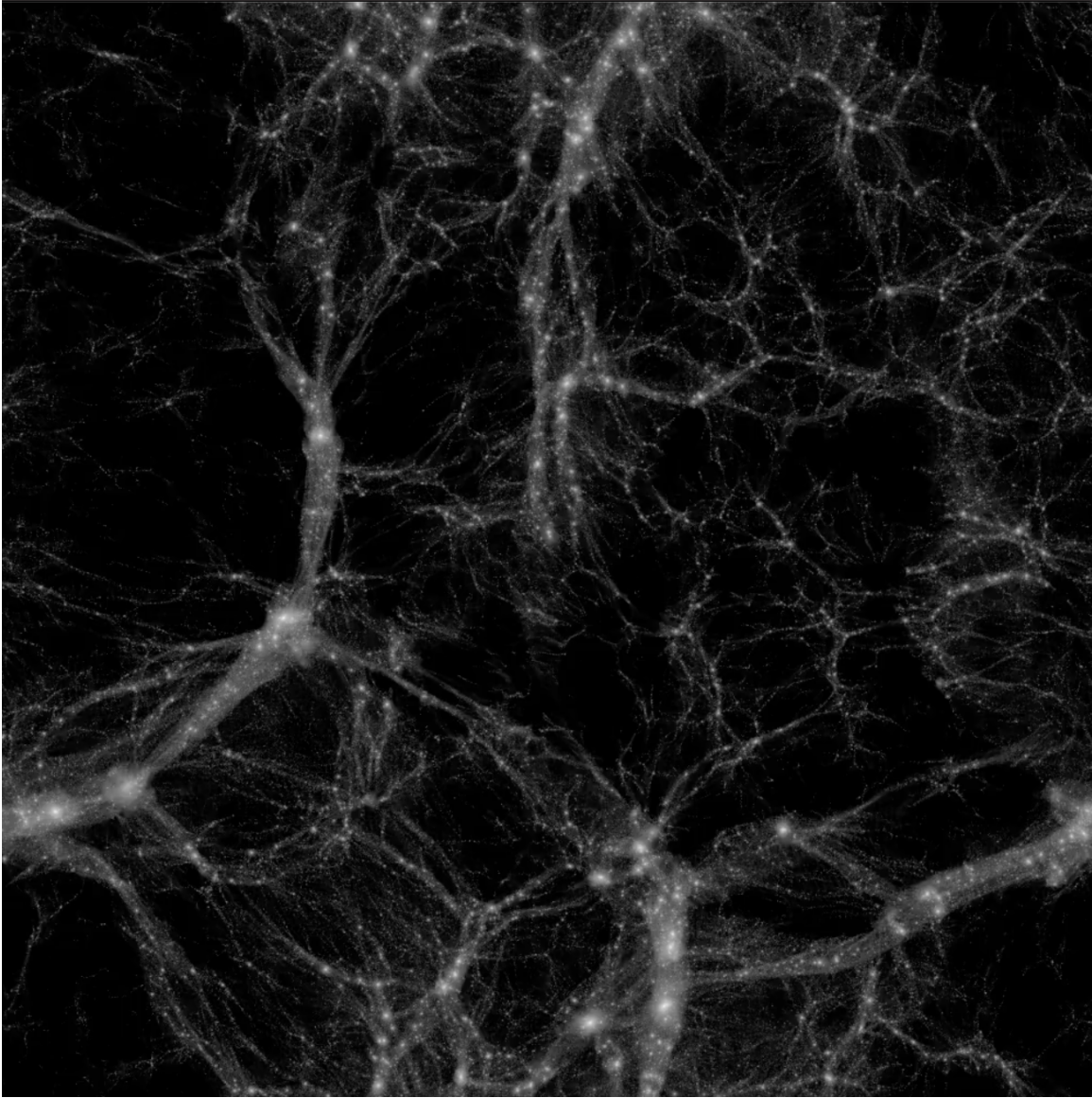
Figure 1.3: An example of the large-scale structure ('the cosmic web') created by dark matter particles (shown as white spots on a black background). Galaxies form from the baryons that gather in the gravitational wells created by dark matter once. This is a thin slice through one of the 100^3 h^{-3} cMpc^3 CAMELS-SAM simulations (Perez *et al.*, 2022).
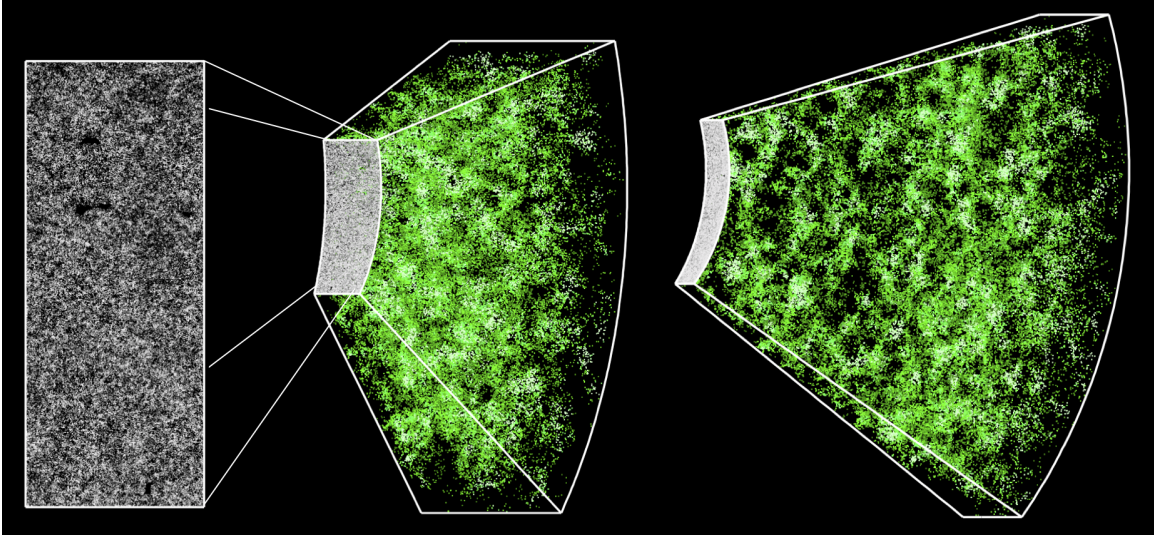
Figure 1.4: An example of the *measured* large-scale structure of galaxies, from the Baryon Oscillation Spectroscopic Survey (BOSS) program of Sloan Digital Sky Survey III. The left-most box covers 1000 sq. degrees in the sky and contains ∼120,000 galaxies, for ∼ 10% of the total survey. Using spectroscopically-measured redshifts reveals the 3D structure to approximately $z < 0.8$, or 7 billion years into the past. *https://www.sdss.org/press-releases/astronomers-map-a-record-breaking-1-2-million-galaxies-to-study-the-properties-of-dark-energy/*

In this section, I introduce those we will encounter throughout this work.

1.1.1   The Two-Point Correlation Function

The most common of the statistics we will encounter is the widely used two-point correlation function (2ptCF). The 2ptCF robustly quantifies the likelihood of two galaxies existing within a certain distance from each other compared to random, at the cost of various types of pair counts across the data and a large random sample.

The 2ptCF comes in many flavors:

- 3D space: either *real* space, the true transverse comoving positions of galaxies

in a volume; or *redshift* space, where galaxies' measured redshifts are directly translated into comoving positions without accounting for e.g. the 'finger of god' effect (galaxies close together will have large random velocities relevant to each other, leading to elongated over-dense regions in redshift maps)

- 2D angular space: a.k.a. the *angular correlation function $\omega(\theta)$*, used for galaxies observed in narrow redshift ranges

- 2D projected space: a.k.a. the *projected correlation function $\omega_p(r, \pi)$*, which projects galaxies across a larger redshift ranges onto two dimensions, and is common in photometric surveys and for other specific analyses

The root of all these forms is the fundamental 2ptCF in 3D real space, $\xi(r)$. For a randomly selected object in a sample, $\xi(r)$ can be defined through the probability of finding a neighbor at a distance r in volume element δV is (Peebles, 1980a):

$$\delta P = n \delta V \left[1 + \xi(r) \right] \tag{1.1}$$

The Fourier Transform of $\xi(r)$ gives the power spectrum, which by definition quantifies the large scale structure born from small fluctuations in the matter field (as e.g. encoded within the cosmic microwave background). The power spectrum is a crucial measurement for confirming our existing models of cosmology and galaxy formation; for example, the 'bump' in the power spectrum near $R \sim 100$ cMpc or $0.1 < k < 1$ h Mpc^{-1} is known the the baryon acoustic oscillation signal, caused by acoustic density waves in the pre-recombination universe.

Additionally, the 2ptCF acts as a power law in log-log space, which is known for most types of galaxies to have the particular slope of -1.8 at medium to large scales. The amplitude of the power law quantifies the amount of clustering, and yields the *correlation length, r_0*, defined as the scale length of clustering when $\xi = 1$:

$$\xi = \left(\frac{r_0}{r}\right)^{1+\delta} \approx \left(\frac{r_0}{r}\right)^{1.8} \tag{1.2}$$

s

The 2ptCFs are measured in different ways, most often with an estimator such as the Landy and Szalay (1993) estimator:

$$\xi_{\text{L-S}}(r) = \frac{DD(r) - 2DR(r) + RR(r)}{RR(r)} \tag{1.3}$$

The 2ptCF can also be separated into a one-halo and two-halo term. The two-halo term, quantifying the interaction between galaxies in separate halos, makes up the well-behaved linear relationship that yields the correlation function. The one-halo term encodes the interactions between galaxies in the same halo in a bump at smaller distance scales. The one-halo term is especially important for theorists who attempt to connect dark matter halo structure to the observable galaxy clustering (often under the heading of halo occupation distribution, or HOD, modelling).

1.1.2   Count-in-Cells and the VPF

Count-in-cells (CiC) quantifies the number of galaxies within a randomly placed cell of a given size, and theoretically includes all volume-averaged correlation functions. However, it is the most computationally costly of the statistics introduced so far: the 2ptCF counts distances within pairs, and the VPF considers only empty spheres, while CiC must track the number of galaxies in every single sphere However, the investment in CiC has yielded interesting results across diverse works. Uhlemann *et al.* (2020) used CiC clustering to disentangle the effects of massive neutrino mass and σ_8 in the Quijote simulation suite (Villaescusa-Navarro *et al.*, 2020).

The VPF is an uncommon clustering statistic that simply asks: how likely is a sphere of a given size to be empty in a given galaxy sample? This contrasts the widely

used 2ptCF, which quantifies the likelihood of two galaxies existing within a certain distance from each other compared to a random distribution, at the cost of various types of pair counts across the data and a large random sample (e.g. Peebles 1980b; Landy and Szalay 1993). CiC, in turn, quantifies the number of galaxies within a randomly placed cell of a given size and theoretically includes all higher-order n-point clustering statistics (Peebles, 1980b), at the cost of being the most computationally costly of the three statistics discussed. The VPF is simple and efficient to calculate, is tied to all higher order correlation functions as the 0th moment of count-in-cells, and encodes information from higher order clustering that the robust 2ptCF cannot always capture (see: Perez *et al.* 2021; Conroy *et al.* 2005; White 1979).

1.1.3   Cosmic Voids?

The word 'void' has been used to refer to several things in astronomy, and warrants clarification here. There are what are called *cosmic* voids: these are large under-dense regions in the cosmic web that require large detailed sky surveys to map and catalog. They are fascinating sub-structures in our universe, that can have complex shapes (often found with interesting tessellation algorithms: e.g. Sutter *et al.* 2015). Cosmic voids also contain rare and interesting galaxies (e.g. Habouzit *et al.*, 2020), and are being used as alternate cosmological probes (Pisani *et al.*, 2015; Hamaus *et al.*, 2016; Pisani *et al.*, 2019; Zhang *et al.*, 2020; Kreisch *et al.*, 2021).

On the other hand, the 'Void' in the VPF refers to dropped test circles or spheres in a sample that contain no galaxies. They are imagined, randomly placed cells in a sample, and are not considered a 'void' unless they have zero galaxies. (Though the *under*-density probability function, considering cells with fewer than some limit of galaxies, has been used in similar ways, e.g. Tinker *et al.* 2008) The VPF will sense the presence of cosmic voids, but cannot classify or detect them specifically. In the

realm of cosmic voids, some specialized statistical tools have been proposed, such as the void mass function (e.g. Bayer *et al.* 2021).

1.2   The Chapters of this Dissertation and Their Roles

Next, I explain how each chapter of this dissertation expands on how the VPF is used for studying the clustering of LAEs.

1.2.1   Chapter II: The Basics of the VPF & Hierarchical Scaling

Chapter II (in the literature as Perez *et al.* (2021)) sets the stage for how the VPF is and should be measured. As the first project and publication I ever worked on, much of my growth as an early career scientist lives between its lines.

In the context of this dissertation, it can be thought of as where I learned to wield the VPF 'hammer', and test it out on simulated LAEs and on the uncommon application of hierarchical scaling. This chapter measures in great detail the VPF and 2ptCF of a medium-sized set of simulations of LAEs, with a detailed discussion of how the VPF's errors should be quantified. The most widespread and useful result is arguably our derived straight-forward and logical guidelines for when the VPF should be measured, given a sample's density, volume, and desired VPF precision. The VPF 'hammer', the intuition around its uncertainty, the guidelines for survey construction, and the creativity to apply it where not always expected, are the beating heart that carries throughout the later more complex (and perhaps more broadly interesting) chapters.

This work also explored a fundamental difference between the 2ptCF and the VPF/CiC: what it means for a correlation function to be *volume-averaged*. The 2ptCF is able to have a one-halo term precisely because it does not measure a volume-averaged statistic: it uses the distances between every single galaxy and every single

random point dropped among them to approximate the excess probability of a galaxy existing at a particular distance scale. The VPF and CiC are measured instead by counting cells with a given number of galaxies, meaning each distance scale is averaged across the entire volume the sample encompasses. The VPF is powerful because it is related to all (volume-averaged) higher-order correlation functions through *hierarchical scaling.* Galaxies appear to follow a negative binomial model for this hierarchical scaling between the volume-averaged correlation functions. (This means, unfortunately, that the VPF cannot give a shortcut to measuring $\xi(r)$, but can measure instead its volume average, $\bar{\xi}(r)$.)

1.2.2 Chapter III: Probing Reionization with the VPF of LAEs in LAGER

With the VPF 'hammer' prepped, and an initial familiarity with the particularities of simulated LAEs, we shift to the core question of this dissertation: constraining reionization with the VPF of LAEs. A paper based on the work of this Chapter has been submitted to the Astrophysical Journal.

Several works have used the angular two-point correlation function for reionization constraints (ACF; e.g. Ouchi *et al.* 2010; Santos *et al.* 2016), yet few have leveraged the VPF, perhaps better suited to notice the neutral patches typical of inhomogeneous reionization (McQuinn *et al.* 2007; Gangolli *et al.* 2021). This Chapter, and the next, takes the VPF as it is–a slice of the full CiC, with errors and guidelines carried plainly on its sleeve–and show how it might be used practically to constrain reionization in either a narrowband survey at $z = 6.9$. This Chapter introduces a way to additionally assess the VPF in an uncommon way: by sticking to select distance scales, in order to most overtly distinguish different ionization fractions while also considering the where the VPF can be measured for these simulations. We present the robust constraints the VPF can give for reionization in a narrowband survey under this analysis, as well

as guidelines for reionization-era survey construction.

1.2.3   Chapter IV: Constraining Reionization with the VPF of LAEs from Roman

In Chapter 4, we shift to consider the 'next big thing' for extragalactic science: the launch and observation of the Nancy Grace Roman Space Telescope in the next few years. With a mirror as sensitive as Hubble's, a wide-field instrument capturing 200 times WFC3's field of view, and sensitive grism and prism elements that will observe LAEs from $6 < z < 14$, Roman will answer long-lasting questions about the sources, timing, and pacing of reionization (among so much more!).

In this work, we leverage the simulations and lessons of Chapter III, and specifically consider: what can the VPF of Roman LAEs reveal about reionization? First, we transform the reionization-era LAE catalogs of Jensen *et al.* (2014) in two ways: we project the catalogs for different types of reionization (fast vs. slow? Early vs. late?), and then apply Roman-specific observational selections. Then, we create sub-slices that correspond to 1, 4, or 16 deg^2, and measure the VPF of these Roman-like LAE samples. We quantify the constraints each of these areas reach, and clarify just what the VPF and Roman are capable of learning about reionization together. A paper based on this work has been submitted to the Astrophysical Journal.

1.2.4   Conclusion and Appendices

I conclude with brief reflections on what all this work has taught me and my collaborators, and the future of the VPF among LAEs, reionization, and beyond. The appendices expand on ancillary concepts, or additional explanatory figures for VPF distributions.

Chapter 2

VOID PROBABILITY FUNCTION OF SIMULATED SURVEYS OF HIGH-REDSHIFT LYMAN-ALPHA EMITTERS

Lucia A. Perez[1], Sangeeta Malhotra[2], James E. Rhoads[2], and Vithal Tilvi[1]

[1] Arizona State University School of Earth and Space Exploration, Tempe, AZ

[2] NASA Goddard Space Flight Center, Greenbelt, MD

ABSTRACT

We calculate the void probability function (VPF) in simulations of Lyman-α emitters (LAEs) across a wide redshift range ($z = 3.1,\ 4.5,\ 5.7,\ 6.6$). The VPF measures the zero-point correlation function (i.e. places devoid of galaxies) and naturally connects to higher order correlation functions while being computationally simple to calculate. We explore the Poissonian and systematic errors on the VPF, specify its accuracy as a function of average source density and the volume probed, and provide the appropriate size scales to measure the VPF. At small radii the accuracy of the VPF is limited by galaxy density, while at large radii the VPF is limited by the number of independent volumes probed. We also offer guidelines for understanding and quantifying the error in the VPF. We approximate the error in the VPF by using independent sub-volumes of the catalogs, after finding that jackknife statistics underestimate the uncertainty. We use the VPF to probe the strength of higher order correlation functions by measuring and examining the hierarchical scaling between the correlation functions using count-in-cells. The negative binomial model (NBM) has been shown to best describe the scaling between the two point correlation function and VPF for low-redshift galaxy observations. We further test the fit of the NBM by directly deriving the volume averaged two-point correlation function from the VPF and vice versa. We find the NBM best describes the $z = 3.1, 4.5, 5.7$ simulated LAEs, with a 1σ deviation from the model in the $z = 6.6$ catalog. This suggests that LAEs show higher order clustering terms similar to those of normal low redshift galaxies.

2.1 Introduction

Galaxies are clustered in space according to the clustering of dark matter halos and the bias introduced by the baryonic physics involved in galaxy formation, both of which evolve over cosmic time (Bernardeau *et al.* 2002; Benson 2010; Coil 2013). The clustering of galaxies is is commonly studied with the two-point correlation function, which describes the probability of finding a galaxy within a given distance of another. The clustering can then be used to measure the power spectrum and constrain cosmological parameters. In observations, the two-point correlation function is used to measure the correlation length, which allows the comparison of clustering across galaxy samples and how galaxy clustering traces that of the dark matter halos. The clustering of galaxies not only depends on their relationships to their host dark matter halos and the growth of structure under gravity, but also is related in complicated ways to galaxy properties (like color and luminosity) and their local environment (eg. Zehavi *et al.* 2002; Croton *et al.* 2005; Cooper *et al.* 2007; Skibba *et al.* 2014). Additionally, the volume-averaged correlation functions can be compared to perturbation theory and theoretical gravitational statistics (Peebles, 1980a), and are measured using the statistical moments of the galaxy count-in-cells distribution. While the standard two-point correlation function is commonly used in the literature, it does not fully capture details of structure such as filaments and voids (Maurogordato and Lachieze-Rey, 1987) and does not give information about higher order clustering correlations.

The void probability function (VPF) is a less common clustering measurement that describes the probability that a sphere of a given size will contain no galaxies. Sometimes called the 'zero-point' correlation function (the average distance where no galaxies exist), the VPF ties theoretically to the higher order correlation functions

(White, 1979). The count-in-cells statistic includes the VPF and information about several averaged correlation function moments, and appears to follow a predictable pattern of scaling due to gravity. The VPF and two-point and higher order correlation functions are theoretically connected through this 'hierarchical scaling', the idea that the first galaxies to form trace the first structures to collapse, following the evolution of Gaussian density fluctuations due to gravitational instabilities (Coil, 2013). In this framework, the VPF can connect to the higher order correlation functions under an analytic model for the scaling coefficients. The hierarchical scaling between the count-in-cells measured VPF and volume-averaged correlation functions has been found to follow the negative binomial model (NBM) in $z < 1$ galaxy surveys (e.g., Croton *et al.* 2004a; Conroy *et al.* 2005; Tinker *et al.* 2008) and in simulations (e.g., Andrew *et al.* 2013). Conroy *et al.* (2005) deduce and Fry and Colombi (2013) confirm that the NBM of clustering scaling is a feature of the underlying dark matter halos. Additionally, the VPF alone has been used to constrain cosmological parameters of large scale structure (ex. Fry 1986; Otto *et al.* 1986; Maurogordato and Lachieze-Rey 1987; Fry 1988; Little and Weinberg 1994). For example, the VPF was found to be sensitive to Halo Occupancy Distribution (HOD) models with different minimum host masses where the two-point angular correlation function was not (Tinker *et al.* 2006; Berlind and Weinberg 2002) and was able to discern the influence of galaxy assembly bias on HOD modeling (Walsh and Tinker 2019; Beltz-Mohrmann *et al.* 2020).

To study the evolution of clustering across cosmic time, it is important to have a well-defined sample of galaxies that can be imaged over various redshifts. Lyman-α emitters (LAEs) have been observed in large numbers at redshifts $2 < z < 7$ through narrowband surveys (Rhoads *et al.* 2000; Taniguchi *et al.* 2005; Shimasaku *et al.* 2006; Matthee *et al.* 2014; Santos *et al.* 2016; Zheng *et al.* 2016; Ouchi *et al.* 2018; Sobral *et al.* 2018). Lyman-α is the strongest emission signature of the earliest galaxies, likely

coming from active star formation (Partridge and Peebles 1967; Malhotra and Rhoads 2002; Atek *et al.* 2014). Additionally, because many LAE properties do not seem to change significantly over time (Malhotra *et al.*, 2012), they are excellent probes of the evolution of large scale structure and the process of reionization, whether by examining changes to their luminosity functions or clustering signatures brought about by neutral hydrogen attenuating the emission (Malhotra and Rhoads 2004a; Mesinger and Furlanetto 2008; Tilvi *et al.* 2014). Understanding the clustering properties of LAEs and how they have evolved improves our understanding of galaxy evolution, the process and pace of reionization, and how galaxies are tied to their dark matter halos and environments. To date, several works have investigated the clustering properties of LAEs at various redshifts between $z \approx 2 - 7$ with mostly the angular two-point correlation function (e.g. Ouchi *et al.* 2003; Shimasaku *et al.* 2004; Kashikawa *et al.* 2006; Gawiser *et al.* 2007; Kovač *et al.* 2007; Murayama *et al.* 2007; Ouchi *et al.* 2008; Shioya *et al.* 2009; Ouchi *et al.* 2010; Guaita *et al.* 2010; Kusakabe *et al.* 2018; Sobacchi and Mesinger 2015a; Bielby *et al.* 2016; Hao *et al.* 2018; Ouchi *et al.* 2018; Khostovan *et al.* 2018b; Hong *et al.* 2019). However, only Palunas *et al.* (2004) (34 LAEs at $z = 2.38$), Kashikawa *et al.* (2006) (58 LAEs at $z = 6.5$), McQuinn *et al.* (2007) (200 simulated LAEs at z=7.5 in partially neutral IGM) have used the VPF to measure the clustering of LAEs in the literature.

As seen in this sampling of the greater field of galaxy clustering research, the VPF has been used in applications that do not often overlap or inform each other. The VPF has spanned diverse applications, whether used to detect slight signals of clustering for very small galaxy samples (eg. LAEs in Palunas *et al.* 2004; Kashikawa *et al.* 2006); to test hierarchical scaling in large spectroscopic samples of low redshift galaxies (eg. Conroy *et al.* 2005; Croton *et al.* 2004a); or as an additional constraint for theoretical models (eg. Tinker *et al.* 2006; Berlind and Weinberg 2002; Walsh

and Tinker 2019; Beltz-Mohrmann *et al.* 2020). In this work, we seek to clarify when and how to best use it in some of these applications. On what distance scales is a VPF measurement reliable? What might be the most honest way to approximate its uncertainty? How informative is the hierarchical scaling application of the VPF with smaller samples of galaxies? Can the VPF give information about higher-order clustering for high redshift starburst galaxies?In local galaxies, the scaling behavior of higher order correlation functions has helped inform the influence of the local environment on galaxy clustering (Baugh *et al.*, 2004), which is to be expected as galaxy properties are influenced by their environment (eg. Croton *et al.* 2005) and galaxy clustering and higher order correlation terms change with those same properties (eg. Croton *et al.* 2004b; Conroy *et al.* 2005). Can the VPF give similar insight for smaller samples of high redshift galaxies? Finally, how can these insights into the VPF inform future clustering studies within these frameworks and others?

In this work, we analyze simulations of LAE-inhabited dark matter halos across a wide redshift range (Tilvi *et al.*, 2009) to test the limits of the VPF. We measure the hierarchical scaling of the LAEs with count-in-cells, and test the fit by moving between the volume-averaged two-point correlation function and VPF in both directions. Our VPF measurements of the simulated LAE catalogs of Tilvi *et al.* (2009) can serve as comparisons to future clustering studies both in their measurements and how our uncertainty is calculated. We confirm that higher redshift LAEs are more clustered and their hierarchical scaling follows the negative binomial model to at least $z = 6$. This could indicate that, under the T09 model, Ly-α emission does not show strong higher order clustering, though the scale of our simulations might not be enough to detect this signal.Finally, the good agreement between the VPF and other clustering measurements we find shows that the VPF could replace the count-in-cells method in contexts that do not specifically probe hierarchical scaling (such as Jensen *et al.* 2014

do to detect the influence of inhomogeneous reionization on the clustering of LAEs), while carrying through more intuitive errors that reflect the properties of the galaxy observations.

This paper is organized as follows. In §2.2, we give the relevant details of the Tilvi *et al.* (2009) LAE simulations. In §2.3, we define the VPF, explore its uncertainty and error, and offer guidelines for its use. In §2.4, we measure the volume-averaged and standard two-point correlation function, contextualize hierarchical scaling, and motivate the use of the negative binomial model. In §2.5, we examine the hierarchical scaling behavior of our simulated LAEs. We measure the hierarchical scaling of our simulated LAEs with count-in-cells to show the negative binomial model is the best fit. We test the negative binomial model against our catalogs by deriving a VPF curve from the volume-averaged correlation function, as did Conroy *et al.* 2005 (hereafter C05) and Croton *et al.* 2004a (hereafter Cr04). Then, we invert the transformation to derive the correlation function for these catalogs from their VPFs as an additional test of the negative binomial model. We give our main conclusions and final remarks in in §2.6.

2.2   Simulated Lyman-α Emitter Catalogs

Tilvi *et al.* 2009 (hereafter T09) presented a simple model of populating dark matter halos with Lyman-α emitters using large cosmological simulations. Here we briefly summarize relevant details and results. Their model assumes that Lyα luminosity is proportional to the star formation rate (SFR), which is directly related to the mass accreted onto the host dark matter halo in the last 30 Myrs.

T09 generated catalogs of dark matter halos using the N-body ΛCDM cosmological simulation GADGET2 (Springel, 2005) with initial conditions from second-order Lagrangian perturbation theory (Crocce *et al.* 2006; Thacker and Couchman

2006). The simulations contained $(1024)^3$ dark matter (DM) particles of mass $2.7 \times 10^7 \ M_\odot \ h^{-1}$ in a volume of $(102 \ \mathrm{cMpc})^3$ or $(73 \ h^{-1} \ \mathrm{cMpc})^3$. Using a friends-of-friends halo finder (Davis *et al.*, 1985), they tracked DM halos with at least 100 particles. The simulation was run from $z \approx 10$ to $z \approx 3$ while tracking the positions and masses of halos.

The DM halos with positive accretion rates are given a Lyman-α line luminosity using this model with star formation rate (SFR) and star formation efficiency (SFE):

$$L_{\ \mathrm{Ly}\alpha} \ = \ 1 \times 10^{42} \ \frac{\mathrm{SFR}}{M_\odot \mathrm{yr}^{-1}} \ \mathrm{erg \ s}^{-1} \ ; \ \mathrm{SFR} = \mathrm{SFE} \left(\frac{\Delta M_b}{t_{\mathrm{Ly}\alpha}} \right) \ = \ \mathrm{SFE} \left(\frac{\Omega_b}{\Omega_{\mathrm{DM}}} \right) \left(\frac{\Delta M_{\mathrm{DM}}}{t_{\mathrm{Ly}\alpha}} \right)$$

$$(2.1)$$

Here, ΔM_b is the amount of baryonic mass accreted by the DM halos, and $t_{\mathrm{Ly}\alpha} = 30$ Myr is the short timescale over which this accreted mass is converted into new stars in the model. T09 chose $t_{\mathrm{Ly}\alpha} = 30$ Myr based broadly on the ages of stars in observed LAEs, the lifetime of OB associations, and the dynamical time that the size of LAEs predicts (e.g. Finkelstein *et al.* 2007; Pirzkal *et al.* 2007; Finkelstein *et al.* 2008). The accreted baryonic mass is derived from the universal ratio of baryonic to DM densities (Ω_b, Ω_{DM}), and the mass accreted onto a DM halo at each step of the simulation (ΔM_{DM}). The model assumed that the escape fraction of Lyα is 1 and that of Lyman continuum photons is 0. In observations, the Lyα escape fraction of LAEs has been measured at about 10 to 50 percent (Nakajima *et al.* 2012; Matthee *et al.* 2016; Sobral *et al.* 2018), and possibly increasing with redshift and with complicated relationships to Lyα equivalent width (Sobral *et al.* 2017; Harikane *et al.* 2018; Oyarzún *et al.* 2017; Trainor *et al.* 2019).

Table 2.1: Relevant Details from Tilvi *et al.* (2009) for clustering measurements. The column headers stand for: redshift (at the center of the simulation box); the Lyα line luminosity cut applied; the total number of simulated LAEs in the catalog; the number of LAEs after the luminosity cut; the surface density of galaxies after the luminosity cut in arcmin^{-2} (whole volume, and two halves when cut evenly in redshift space); the volume density of galaxies after the luminosity cut in cMpc^{-3}; and the observations of LAEs to which the catalogs were designed to compare.

z	L_{cut}, 10^{42} erg s^{-1}	N_{total}	N_{cut}	Surface Density $\Sigma_{arcmin^{-2}}$	Volume Density $\mathcal{N}_{Mpc^{-3}}$	Comparison observations
3.1	1.995	62,364	1,145	Whole (102 cMpc)^3: 0.4029	1.079×10^{-3}	Gawiser *et al.* (2007) &
				Front $102 \times 102 \times 51$ cMpc^3: 0.1717		Khostovan *et al.* (2018b)
				Back $102 \times 102 \times 51$ cMpc^3: 0.2312		
4.5	1.673	64,868	1,211	Whole ($0.75 \times (102$ cMpc)^3): 0.5765	1.521×10^{-3}	Kovač *et al.* (2007)
				Front $102 \times 102 \times 51$ cMpc^3: 0.3842		Ouchi *et al.* (2003)
				Back $51 \times 102 \times 51$ cMpc^3: 0.3847		
5.7	3.068	79,429	539	Whole (102 cMpc)^3: 0.3014	5.08×10^{-4}	Ouchi *et al.* (2010) &
				Front $102 \times 102 \times 51$ cMpc^3: 0.1197		Ouchi *et al.* (2008)
				Back $102 \times 102 \times 51$ cMpc^3: 0.1817		
6.6	3.068	79,783	355	Whole (102 cMpc)^3: 0.2171	3.35×10^{-4}	Ouchi *et al.* (2010)
				Front $102 \times 102 \times 51$ cMpc^3: 0.0960		
				Back $102 \times 102 \times 51$ cMpc^3: 0.1211		

In the simple model of T09, LAEs fundamentally act as tracers of DM halo buildup and the accompanying cold gas accretion. T09 apply an analytic description to populate dark matter halos with LAEs based only on the evolution of the N-body simulation. Progress from many angles has occurred in the field of LAE modeling since. Similar studies that worked primarily with smoothed particle hydrodynamic simulations and intrinsic galaxy properties include: Nagamine *et al.* 2010 (who also consider the intrinsic star formation in LAEs, but do not consider dust enrichment or variance in the intergalactic medium transmission); Dayal *et al.* 2009 and Dayal *et al.* 2010 (whose model includes in the luminosity of stellar sources, cooling of HI, dust enrichment, and IGM transmission); and Kobayashi *et al.* 2007 and Kobayashi *et al.* 2010 (who apply the Mitaka semianalytical hierarchical clustering model for star formation from Nagashima and Yoshii 2004 to account for extinction from interstellar dust and outflow feedback). Many works have also included radiative transfer calculations, for example to prepare for epoch of reionization by expanding how their models account for the effect of neutral IGM on the Lyα line (McQuinn *et al.* 2007; Iliev *et al.* 2008; Zheng *et al.* 2010; Dayal *et al.* 2011 Jensen *et al.* 2013; Kakiichi *et al.* 2016; Inoue *et al.* 2018; Gangolli *et al.* 2021; and more.)

The simple model of T09 fits the single parameter of star formation efficiency to find good agreement with LAE observations at redshifts 3.1, 4.5, 5.7, and 6.6. The Lyα luminosity function of the $z = 3.1$ simulation is used to set the SFE by comparing with the LF of the Gronwall *et al.* (2007) LAE observations at the same redshift. With all parameters defined or derived in the model, T09 match the Lyα line luminosity limits of surveys of LAEs at the same redshifts to test the strength their model. They reproduce the Lyα luminosity functions, star formation rates, approximate halo masses, and duty cycles of comparable LAE surveys and observations with a simple adjustable parameter for all redshifts. Finally, their derived correlation lengths track

closely with those observed for LAEs at similar redshifts. These simulations are therefore ideal for this exploration of the VPF for LAEs at varying redshifts.

For this work, we use the output catalogs that give every LAE's position in the $x - y - z$ space (between 0 and 73 h^{-1} cMPc) and their Lyman-α line luminosity. For the z=4.5 catalog, we exclude the back left quadrant of $x < 36$ h^{-1} cMPc and $z > 36$ h^{-1} cMPc (the top and bottom 'back' left (51 cMpc)^3 sub-volumes) because the simulation save file for this redshift was apparently truncated, omitting most DM halos (and therefore LAEs) from this subvolume. The rest of the volume behaved as expected, so we exclude this artificially empty region in our analysis. Accounting for this improves the agreement of the z=4.5 simulated LAEs with the comparison observations of Kovač *et al.* 2007 (hereafter K07) and Ouchi *et al.* 2003 (hereafter O03), and explains the slight inflation of the $z = 4.5$ correlation length in T09.

When plotting our VPF measurements, we convert the transverse comoving positions to transverse comoving megaparsec (hereafter cMpc) using T09's $h = 0.716$, so that the positions range from 0 and 102 cMpc. When calculating all correlation lengths, we revert to units of h^{-1} cMpc to directly compare with earlier measurements of observed LAEs, whose angular separations were converted into physical distances assuming $h = 0.7$ (from $H_0 \equiv 100 \ h$ km s^{-1} Mpc^{-1} = 70 km s^{-1} Mpc^{-1}).

Table 2.1 provides relevant details of the catalogs for this work. **Complete Table 2.1 caption/header descriptions:** The column headers stand for: redshift (at the center of the simulation box); the Lyα line luminosity cut applied to match the catalog to the corresponding comparison observation(s); the total number of simulated LAEs in the catalog; the number of LAEs after the luminosity cut; the surface density of galaxies after the luminosity cut in arcmin^{-2}, for the whole volume and two halves when cut evenly in redshift space; the volume density of galaxies after the luminosity cut in cMpc^{-3}, with each side of the volume measuring 102 cMpc; and the observations

of LAEs to which the catalogs were designed to compare. The 'front' is the area with z-coordinates from 0 to 51 cMpc, and the 'back' is z-coordinates from 51 to 102 cMpc. The $z = 4.5$ catalog has had the 'back' left quarter removed due to a previously undiscovered output error in the simulation.

The luminosity cuts follow the survey limits for observations that each catalog mimicked. We confirm the corresponding Lyman-α line luminosity for the $z = 4.5$ K07 observation and confirm the luminosities that Ouchi *et al.* (2010) found for the observations at $z = 3.1$, 5.7, and 6.6 by using the relevant limiting magnitude and narrowband filter details for each observation. We also compare to the LAE sample of Khostovan *et al.* (2018b) at $z = 3.10$ in the SA22 field using archival narrowband imaging that closely matches the Lyα luminosity limit and volume density of our $z = 3.1$ simulation catalog.

The simulated LAEs of T09 successfully recreated observed luminosity functions, equivalent width distributions, age distributions, and duty cycles. Future Lyα surveys can further test the Tilvi *et al.* (2009) model by a similar comparison of our VPF measurements and hierarchical scaling.

2.3   Void Probability Function

2.3.1   VPF Theory and Algorithm

The Void Probability Function (VPF) is the probability that regions of a particular radius will have no galaxies within them. It contains information about all higher order correlations (White 1979; Maurogordato and Lachieze-Rey 1987), and is the 'zero-point' volume-averaged correlation function. The VPF (labeled P_0 for brevity) of a galaxy sample with mean density n is defined by the hierarchy of all p-point reduced correlation functions w_p at a given volume V:

$$P_0(V) = \exp\Big(\sum_{p=1}^{\infty} \frac{(-n)^p}{p!} \int ... \int w_p(x_1, ..., x_p) \, dV_1...dV_p\Big) \qquad (2.2)$$

The VPF is calculated by simply counting what fraction of randomly placed test spheres are empty as a function of radius. Our algorithm takes in galaxy locations as point sources on an $x - y$ (2D) plane or $x - y - z$ (3D) volume, and for each radius being tested, generates many random central positions within the area (that would not have the radius overlap the boundaries), and counts how many of the test spheres have no galaxies within the radius.

When we generate the randomly placed test spheres, we place ten thousand points to guarantee we sample all true voids. We repeat the VPF measurement for all radii 50 times to account for the small variance that comes when choosing random points and guarantee that we minimize the error from not completely sampling our volumes (hereon out the *sampling error*, which C05 used as the error on their VPF). Throughout this work, we plot errors in relative logarithmic space:

$$Y = \log_{10}(P_0) \quad \rightarrow \quad \delta Y \approx 0.434 \times \frac{\delta(P_0)}{P_0} \qquad (2.3)$$

Figure 2.1 displays the VPF calculations of our four catalogs as colored stars (3D) or colored circles and triangles (2D, 'front' and 'back' halves). The 2D VPF is the most easily implemented tool for observational surveys, since redshift space data are arduous to gather and introduce additional error, while the angular positions of objects in the sky are known with excellent precision. When there exist accurate redshift data, the three-dimensional VPF in redshift space can be explicitly connected to the correlation functions by assuming given hierarchical models (eg. C05). §2.5 details the connection between the three-dimensional VPF and volume-averaged three-dimensional two-point correlation function for these simulations.

Full Figure 2.1 caption: *Left: the measured 2D VPF*. For the 2D VPF, we have split our volumes into 'front' and 'back' halves (split halfway in the z direction) to match the thickness of the 8 independent cubical sub-volumes. The circles/triangles mark the 2D VPF of the front/back halves. The black dashed/dotted lines trace the 2D VPF of unclustered points at the same surface density of the front/back halves. *Right: the measured 3D VPF*. For the 3D VPF, we have marked the theoretical radii where the VPF can be measured based on the number densities, volume, and our chosen precision of $\pm 10^{-2}$. The stars are the 3D VPF of the whole $(102 \text{ cMpc})^3$ volume $(0.75 \times (102 \text{ cMpc})^3$ for the $z = 4.5$ catalog), and the dashed black line is the 3D VPF of unclustered points at the same volume density. The error bars around the black lines correspond to 3 times the sampling error. The minuscule sampling error confirms that the size of our our galaxy samples and test sphere arrays is large enough when compared to our full volume for a precise measurement. The shaded colored regions correspond to 1σ standard error across the independent $(51 \text{ cMpc})^3$ sub-volumes' VPFs. All our catalogs' VPFs differ significantly from their random VPF curves, confirming that they are all clustered.

2.3.2 Verifying Our VPF algorithm

We check the accuracy of our VPF calculation with randomly distributed points. The original equation for the VPF from White (1979) gives its exact value for a Poisson distribution. This theoretical VPF curve of completely unclustered points depends only on the surface or volume density of the sample. With N as the number of galaxies sought in the volume V, and n as the mean particle density, the probability to find N galaxies in a volume of V is:

$$P_N(V) = \frac{(nV)^N}{N!} e^{-nV} \; ; \quad \text{for} \quad N \;=\; 0 \; , \; P_0(V) = e^{-nV}. \tag{2.4}$$
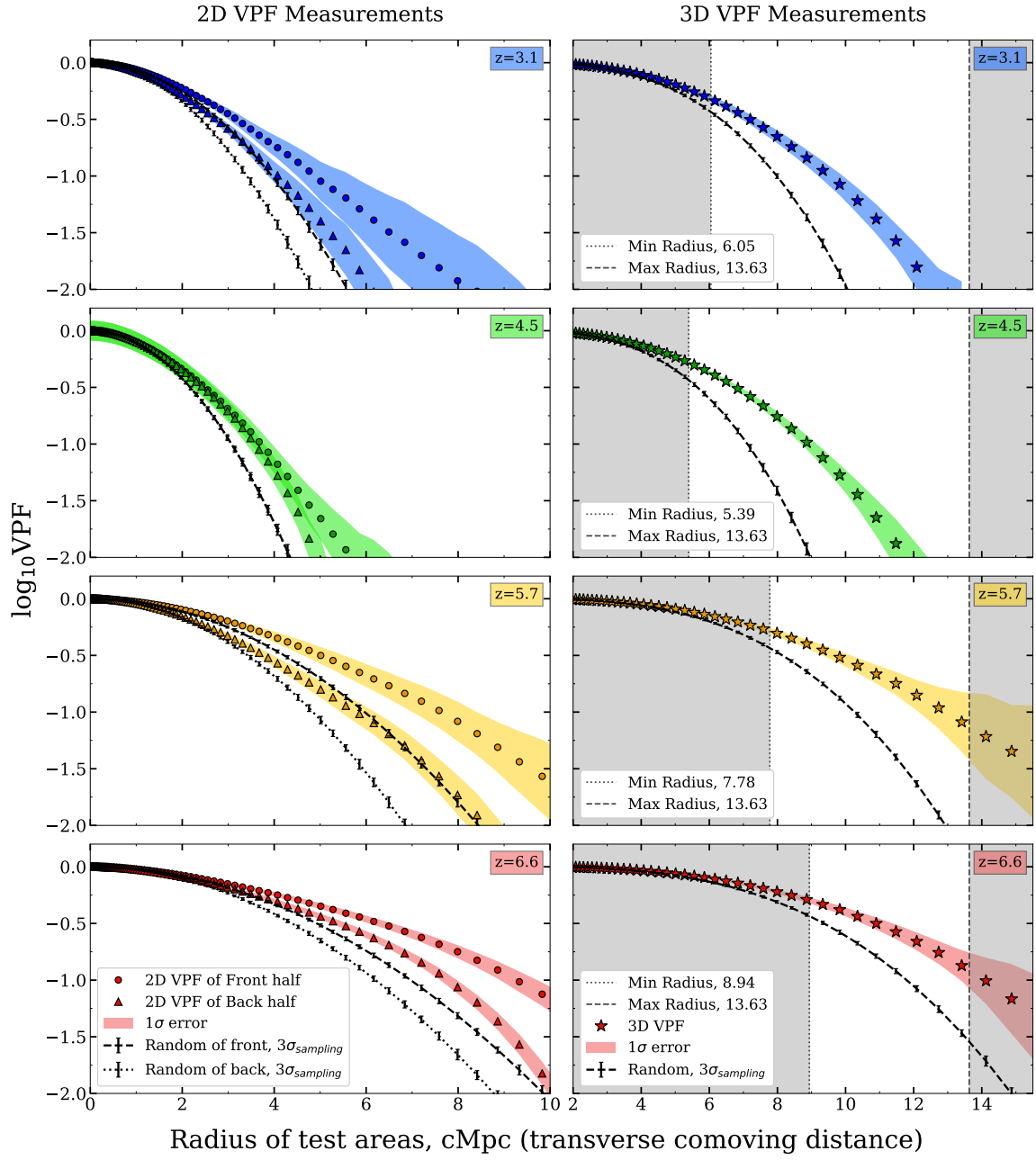
Figure 2.1: 2D (left) and 3D (right) VPF for the four LAE catalogs. Find complete caption at the end of §2.3.3.

We can rewrite this equation to use surface density Σ or volume density $\mathcal{N}$ and the radius of the tested void to yield these expressions for the VPF of completely unclustered points:

$$\log_{10}(P_{0,\text{2D}}(r)) = \frac{-\Sigma \times (\pi R^2)}{\ln(10)} \tag{2.5}$$

$$\log_{10}(P_{0,\text{3D}}(r)) = \frac{-\mathcal{N} \times ((4/3)\pi R^3)}{\ln(10)} \tag{2.6}$$

Deviation from this VPF indicates a sample is clustered. We check our algorithm by generating randomly placed points at the same surface or volume density of the given catalog after the luminosity cuts; measuring their VPF; and finally comparing the curve to the theoretical predictions. We find excellent agreement between the theoretical curves and the measured VPF of randomly placed points, confirming our algorithm is accurately measuring the VPF.

2.3.3   Fundamental Limits of the VPF

Multiple factors contribute to the uncertainty in clustering measurements, some of which have been underestimated in past studies of the VPF. For the VPF, the most fundamental factors are the size and scale of the galaxy sample.

To articulate these dominating constraints to a reliable VPF measurement, we imagine a distribution of N galaxies in a volume of V with number density $\mathcal{N}$, randomly distributed and with no correlation in their positions. The smallest radius to dependably probe the VPF is the volume inhabited by a single galaxy, or the average distance between two galaxies:

$$\mathcal{N}\frac{4\pi}{3}R_{\min}^3 = 1 \quad \rightarrow \quad R_{\min} = \sqrt[3]{\frac{1}{\mathcal{N}}\frac{3}{4\pi}} \tag{2.7}$$

The largest radius to dependably probe the VPF depends instead on the size of the volume and the desired precision of the VPF. The VPF is a fraction often reported as logarithms, so in order to measure $\log_{10}(P_0)$ to a given $-\alpha$ value and guarantee the level of precision of the VPF within $\pm 10^{-\alpha}$, the volume must accommodate 10^α independent sub-volumes of the given radius. For example, if one of these 10^α independent sub-volumes is empty, then the VPF is indeed $1/10^\alpha$. Therefore, the maximum radius to probe VPF is the radius of the independent sub-volumes that probe the desired sensitivity in the sample volume:

$$V = 10^\alpha \frac{4\pi}{3} R_{\text{max}}^3 \quad \rightarrow \quad R_{\text{max}} = \sqrt[3]{\frac{V}{10^\alpha} \frac{3}{4\pi}} \tag{2.8}$$

Combining these minimum and maximum radii measurements allows us to constrain the number of galaxies needed in a sample for the desired measurement. The dynamic range between the radii is then:

$$d = \frac{R_{\text{max}}}{R_{\text{min}}} = \frac{((V/10^\alpha)(3/4\pi))^{1/3}}{((1/\mathcal{N})(3/4\pi))^{1/3}} \quad \rightarrow \quad \sqrt[3]{\frac{\mathcal{N}V}{10^\alpha}} \tag{2.9}$$

A specific dynamic range and VPF sensitivity will constrain a specific number density and volume ratio, and therefore the number of galaxies for the measurement:

$$N_{\text{total}} = \mathcal{N}V = d^3 10^\alpha \tag{2.10}$$

For example, with a dynamic range of 2 in radius (eg. 6 to 12 cMpc) and probing to $\log_{10}(P_0) = -2$, the sample must include at least 800 galaxies. To improve the measurement of the VPF at small scales, we must increase the number density. To improve the measurement at large scales, we must increase the volume.

These derivations rely on assuming a completely unclustered galaxy distribution, and are therefore very conservative guidelines. When galaxy samples display clus-

Table 2.2: Theoretical Radii VPF Limits for Our Catalogs, Assuming $\alpha=2$ and $V=(102 \text{ cMpc})^3$ for a Perfectly Random Distribution.

z	N_{cut}	$\mathcal{N}_{cMpc^{-3}}$	R_{min}, cMpc	d
3.1	1,145	1.079×10^{-3}	6.05	2.25
4.5	1,211	1.521×10^{-3}	5.39	2.53
5.7	539	5.08×10^{-4}	7.78	1.75
6.6	355	3.35×10^{-4}	8.94	1.52

tering, the average distance between galaxies is smaller and the limiting R_{min} will decrease. For our samples, we probe to $\log_{10}(P_0) = -2$. All catalogs have the same underlying $(102 \text{ cMpc})^3$ volumes and same maximum radius of 13.63 cMpc.[1] Table 2.2 shows the minimum radii and dynamic ranges of our catalogs after applying the luminosity cuts. Based on the number of galaxies left after the luminosity cut and the resulting number density of the sample, these are the minimum radii and largest dynamical range of radii over which the VPF can be measured.

In Figure 2.1, we indicate these radii limits for our 3D VPFs with grey shading. In §2.3.3, we restrict ourselves to these radii when we convert our catalog's VPFs into correlation functions. The luminosity cuts we used so we could directly compare to observations decrease the number density and increase R_{min} significantly. Upon examination of Figure 2.1, the limiting radii in Table 2.2 correspond to when the random and VPF curves diverge significantly, and when the error begins to dominate.

---

[1]Although we chose remove 25% of the volume in the $z = 4.5$ simulation to account for a previously undiscovered output error, we maintain this maximum radius for consistency. This choice makes little difference in our final result.

Though we focus on the 3D VPF for the rest of this work, we can also derive these radii limits for the 2D VPF using the survey area A and surface density Σ:

$$\Sigma \times \pi R_{\min}^2 = 1 \quad \rightarrow \quad R_{\min} = \sqrt{\frac{1}{\Sigma\pi}} \tag{2.11}$$

$$A = 10^\alpha \times \pi R_{\max}^2 \quad \rightarrow \quad R_{\max} = \sqrt{\frac{A}{\pi 10^\alpha}} \tag{2.12}$$

$$d = \frac{R_{\max}}{R_{\min}} = \sqrt{\frac{\Sigma A}{10^\alpha}} \tag{2.13}$$

$$N_{\text{total}} = \Sigma \times A = d^2 10^\alpha \tag{2.14}$$

2.3.4 Traditional Statistical Estimators for Error in the VPF

The majority of past studies utilizing the VPF have approximated the error in their measurements by utilizing traditional methods of 'internal' error estimates or creating mock catalogs and measurements ('external' estimates). 'Internal' error estimates cut the entire sample into smaller sections and repeat the statistical measurement in different iterations. Norberg *et al.* (2009) give an excellent review and assessment of internal error estimation methods for correlation function clustering studies. They determine that the bootstrap method with oversampled sub-volumes agrees best with their external error estimation methods (galaxy formation models with fully known inputs). In the *bootstrap* method, a resampling of the data set is made by randomly selecting some number of sub-volumes while allowing for replacement and repetition.

However, conclusions for the two-point correlation function error behavior will not exactly translate to the VPF. Past studies using the VPF observed galaxies often use

the *jackknife* method, where the volume is divided into subsections, each subsection is systematically left out the full volume for a measurement, and the mean and variance of the final distribution is taken to approximate the mean and variance of the original. For example, C05 exclusively use the jackknife technique to estimate their errors on their VPF. Cr04 quote 1σ errors on their VPFs derived from the *rms* scatter over their many generated mock catalogs, and verify that the error from jackknife technique is comparable. Finally, Khostovan *et al.* (2018a) found that the Poisson estimate of the correlation function's error is very consistent with the bootstrapping errors for samples with 10^{2-3} galaxies.

We measure the variance with the traditional jackknife technique with the eight sub-volumes and find it underestimates the error compared to the subsample method we introduce in §2.3.5 by at least a factor of 2 at all scales. Yang and Saslaw (2011) suggested that the jackknife method of C05 underestimated the true range of the data's variability, as subsets of the data are not identically distributed. Therefore, we caution against relying solely on this technique to measure the error on the VPF measurement, while recognizing that more work remains to be done studying how the jackknife technique performs on the VPF of larger samples.

2.3.5 Approximating Error in the VPF with Independent Sub-Volumes

We choose the subsample method to quantify our error in the most 'independent' way possible. Like other gauges of clustering, individual measurements of the VPF are not independent, since clustering also takes place on scales much larger than most observations. In a practical sense, a single abnormally empty or overdense region affects the entire VPF curve for the whole volume. To mimic how truly independent measurements of the VPF within our radii limits would behave, we divide our catalog volumes into the largest equally sized independent subvolumes that they hold (72

31

cMpc h^{-1} ÷ 2 = 36 cMpc h^{-1} or 51 cMpc per side), and use the spread in the subvolumes' VPF measurements to guide the error in our whole volume's VPF. Though such subdivisions of galaxy data are not truly independent, this method helps illuminate the inherent variance of the VPF at radii less than 15 Mpc.

In Figure 2.1, we shade the subsample error around the VPFs of the 'halves' (2D) and the whole volume (3D) in the corresponding colors. For this method, we divide our simulated catalogs into eight non-overlapping independent sub-volumes of $(51 \text{ cMpc})^3$. (The $z = 4.5$ catalog ends up with six non-overlapping independent sub-volumes, due to the apparent output error in the back left corner that was not discovered until this work.) We measure the VPF ten times for each of the sub-volumes with 5000 randomly dropped test voids every time. For an approximation of the error of the whole volumes' VPF at each radius, we measure the standard error in the population mean (the standard deviation of the 80 VPF measurements across all sub-volumes and iterations divided by the square root of 8):

$$1\sigma \text{ standard error } = \frac{\text{STD}(P_{0,\text{subvols}})}{\sqrt{N_{\text{subvols}}}} \quad \rightarrow \quad \frac{\sqrt{|P_{0,i} - \bar{P}_0|^2}}{\sqrt{8}} \tag{2.15}$$

2.4 Hierarchical Scaling and the Correlation Functions

2.4.1 Measured Two-Point Correlation Functions & Correlation Lengths

The two-point correlation function measures the excess probability of finding a galaxy at a given separation r from another, relative to a random Poisson distribution (Peebles 1980a; Coil 2013):

$$dP = n[1 + \xi(r)]dV, \tag{2.16}$$

where n is the mean number density of the sample. The two-point correlation

function is measured in three dimensions in comoving space, and yields the power spectrum with a Fourier Transform (see Peebles 1980a for a thorough text on large scale structure). We use the three dimensional form of the two-point correlation function estimator from Landy and Szalay (1993), where the data and randomly distributed point catalogs are the same size:

$$\xi_{\text{L-S}}(r) = \frac{DD(r) - 2DR(r) + RR(r)}{RR(r)} \tag{2.17}$$

The two-point correlation function of galaxy samples is known to mostly follow a power law with a slope of $1+\delta$, with the long-used value of $\delta=0.77 \pm .006$ from Peebles (1975) and Totsuji and Kihara (1969). The amplitude of the power law gives the correlation length r_0, which is defined as the scale length of clustering for a given galaxy sample when $\xi = 1$. The correlation length is often expressed in terms of h^{-1} cMpc to facilitate comparison between observations with different cosmological assumptions. A larger correlation length value roughly indicates more clustering. To calculate the best fit correlation length for each catalog, we fit this power law on the $\log_{10}(r)$ (h^{-1} cMpc) vs. $\log_{10}(\xi_{\text{L-S}}(r))$ plot for the two-point correlation function:

$$\xi = \left(\frac{r_0}{r}\right)^{1+\delta} \approx \left(\frac{r_0}{r}\right)^{1.8} \tag{2.18}$$

T09 previously calculated the two-point correlation functions and measured correlation lengths for their simulated LAEs, but to maintain transparency and account for small differences in luminosity cuts in our analysis, we repeat the measurements here. We verify the accuracy of the calculation and correlation length fit by finding great agreement with T09's correlation lengths for the $z = 3.1, 5.7, 6.6$ catalogs. Due to the previously undetected output error in the $z = 4.5$ catalog in T09, we ignore the affected corner and measure the clustering the rest of the volume. We compare

directly to the observed LAEs of K07 and O03 and find much better agreement with their correlation length measurements.

First, we calculate each catalog's 'data-data' distance counts ($DD(r)$): for each data point, we measure three-dimensional distance $d = \sqrt{x^2 + y^2 + z^2}$ to every other data point and organized each distance into the corresponding radius bin. Then, for the DR and RR terms, we generate random points in the same way as the VPF random circles were (but here, matching the number of galaxies). For each random point, we compare distance to other random points (creating the $RR(r)$ array) and to the data points (creating the $DR(r)$ array). We repeat the DR and RR tallying process with 1000 different random point sets, as to eventually have 1000 unique $\xi_{\text{L-S}}(r)$ measurements for each catalog.

In Figure 2.2, we plot the median $\xi_{\text{L-S}}$ across all 1000 repetitions with traditional Poisson errors: $\Delta\xi = (1 + \xi)/\sqrt{DD}$. This is an irreducible error in the correlation function. As Khostovan $et\ al.$ (2018a) found for their emission line galaxy samples, the Poisson error is consistent with the computationally expensive bootstrapping technique for galaxy samples between 10^2 and 10^3. Additionally, we plot the volume-averaged two-point correlation function $\bar{\xi}$ we measure in §2.5.1 using count-in-cells in colored circles, along with the Poisson error. The transition to the volume-averaged two-point correlation function from the traditional form (eg. Landy and Szalay 1993) increases the amplitude of the power law by approximately $\log_{10}(3/(2-\beta)$ (regardless of the volume scale, if assuming an ideal power law of slope $1 + \beta$).[2] Because the two-point correlation function is not an ideal power law and drops off at the smallest and largest scales, the $\bar{\xi}$ measurement from CiC will approximate a power law with a less drastic amplitude increase. CiC was shown to be consistent with the volume

---

[2]For those curious about moving from $\bar{\xi}_2$ to the standard ξ_2, see the derivation in Yang and Saslaw (2011).

average of the Landy and Szalay (1993) estimator by Szapudi (1998), and we confirm this in our LAEs.

To best approximate the value and error on the correlation lengths for the samples, we fit each of the 1000 individual $\xi_{\text{L-S}}$ curves to a power law[3] with a fixed slope of $1+\delta=1.8$ and choose the median value of the resulting distribution (Figure 2.3). We add the distribution's 1σ standard deviation in quadrature to the median error out of the least squares fitting algorithm covariance matrix to get our final correlation length 1σ error. In Figure 2.2, the color shaded regions and the values listed in the legends correspond to the median $r_0 \pm 3\sigma$ h^{-1} cMpc. The newest correlation lengths, whose 3σ range is shown as shaded regions and described in the legend of Figure 2.2, agree very well with the T09 calculations and/or those of comparison observations. The new and previous correlation length measurements for $z = 3.1, 5.7, 6.6$ agree within all 1σ errors. For the $z = 4.5$ catalog, the previously undetected empty back corner created a discrepancy between the T09 correlation length and those which K07 and O03 measured in observed LAEs. The new correlation length measurements for $z = 4.5$ agree within all 1σ errors with the O03 and K07 measurements.

Complete Figure 2.2 caption: The spatial 3D two-point correlation function of the LAE catalogs and their best power law fits. We calculate the standard two point correlation function with the Landy and Szalay (1993) estimator 1000 times with new random catalogs, and measure the correlation length of each iteration. We use a least-square method that minimizes the difference between the measured $\xi_{\text{L-S}}(r)$ and the one predicted from the power law in Equation 2.18, assuming a fixed slope of $1+\delta=1.8$. The darker colored triangles indicate the mean $\xi_{\text{L-S}}$ value across the 1000 iterations at each radius. The error bars on the colored triangles are the 1σ logarithmic space Poisson error, $\Delta\xi = (1 + \xi)/\sqrt{DD}$.

---

[3]We use *scipy.optimize.leastsq* (Jones *et al.*, 01)
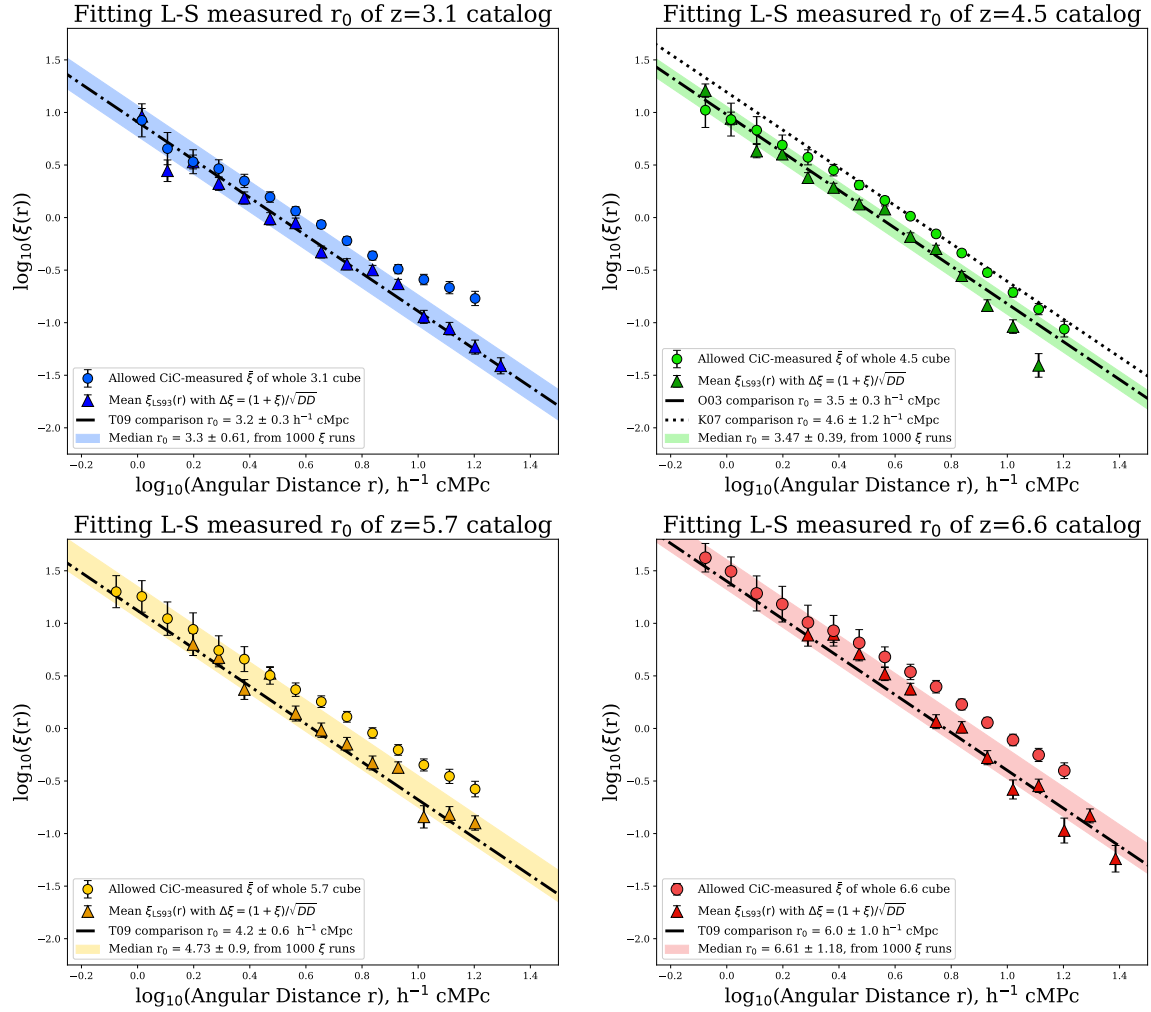
Figure 2.2: The spatial 3D two-point correlation function of the LAE catalogs and their best power law fits. The darker colored triangles indicate the mean $\xi_{\text{L-S}}$ value across the 1000 iterations at each radius. The error bars on the colored triangles are the 1σ logarithmic space Poisson error, $\Delta\xi = (1 + \xi)/\sqrt{DD}$.
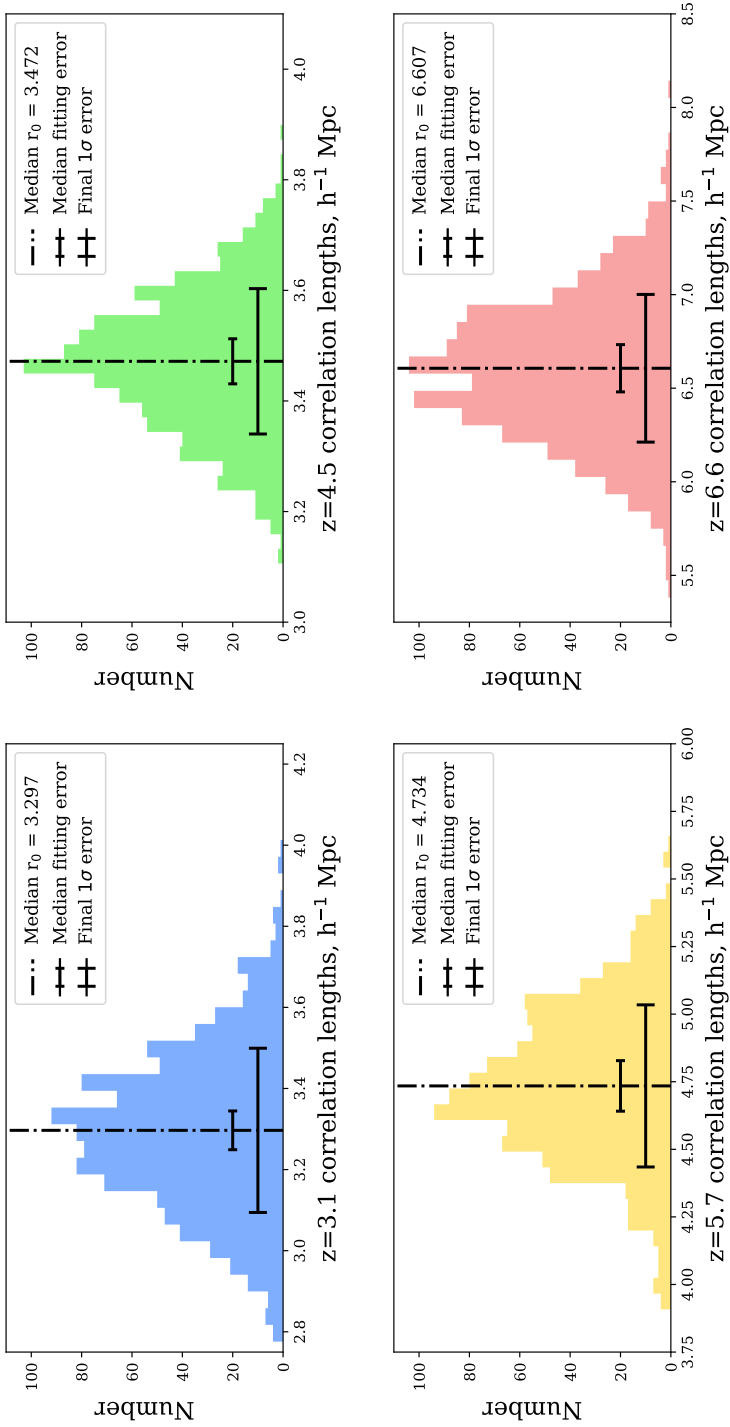
Figure 2.3: Detailed correlation length fits to the LAE catalogs. We calculate the two point correlation function ξ with the Landy and Szalay (1993) estimator 1000 times with new random catalogs, and measure the correlation length of each iteration. The median fitting error from our least squares fitting method is added in quadrature to the 1σ standard deviation across each r_0 distribution to measure a 1σ error on the median correlation length value. The $z = 3.1$, 5.7, 6.6 catalogs agree very well with the previous measurement from T09 and their comparison correlation lengths.

The black dash-dot line marks the best-fit correlation length and error from Tilvi *et al.* (2009) for the $z = 3.1,\ 5.7,\ 6.6$ catalogs, which agree within 1σ errors with our new correlation length measurements. The $z = 4.5$ LAE catalog had an output error undiscovered in T09 that led to excess clustering signals, so we ignore the affected regions when measuring the clustering. We find great agreement within 1σ errors with the correlation length of observed LAEs at $z = 4.5$ from K07 and O03. The final 1σ errors in our correlation lengths come from adding in quadrature the median fitting error across the 1000 calculated correlation lengths and one standard deviation in the correlation length distribution (Figure 2.3). The colored shaded regions and legend entries are the $\pm 3\sigma$ region about the median $r_0\ h^{-1}$ cMpc. The lighter colored circles are the count-in-cells measured volume-averaged two-point correlation function $\bar{\xi}_{\text{CiC}}(r)$, which are used for the VPF transformations later in this work. The higher amplitude is expected with volume-averaged correlation functions, and the pattern of higher clustering at higher redshifts remains.

2.4.2   Hierarchical Scaling and Higher-Order Clustering Moments

Different galaxy distributions can have identical two-point correlation functions but unique VPFs that include additional information from higher order correlation functions (Maurogordato and Lachieze-Rey 1987; White 1979). In the framework of hierarchical scaling, one assumes that all the volume averaged correlation functions are hierarchically related to the volume-averaged two-point correlation function ($\bar{\xi}_2$) via the hierarchical *Ansatz*:

$$\bar{\xi}_p = S_p\ \bar{\xi}_2^{\,p-1}, \quad p \geq 3; \qquad \text{where} \quad \bar{\xi}_p(R) = \frac{\int \xi_p(r) dV_R}{\int dV_R}. \qquad (2.19)$$

This *Ansatz* allows one to transform the VPF of White (1979), now expressing it as a sum of all the higher order volume-averaged correlation functions, their scaling

38

coefficients, and the average number of galaxies in a tested volume of radius R, $\bar{N}(r)$:

$$P_0(r) = \exp\left(\sum_{p=1}^{\infty} \frac{-\bar{N}(r)^p}{p!}\, \bar{\xi}_p(r)\right) \quad \rightarrow \quad P_0(r) = \exp\left(\sum_{p=1}^{\infty} \frac{-\bar{N}(r)^p}{p!}\, S_p\, \bar{\xi}_2^{p-1}(r)\right).$$
(2.20)

One can derive the S_p scaling coefficients by assuming a phenomenological model of hierarchical gravitational clustering (Poisson, Gaussian, negative binomial, thermo-dynamic, etc.). Then, one is able to connect all the correlation functions to the VPF with just the S_p coefficients. This applies only when both are three dimensional and while in redshift space (Kaiser 1987; Ryden and Melott 1996). These phenomenological models predict specific strengths for all the correlation functions, and deviations from these predictions when compared to measured void statistics or higher order correlation functions are used to explore the underlying physics of clustering (eg. Fry *et al.* 1989).

The hierarchical scaling models can connect the VPF to the volume-averaged two-point correlation function and higher-order correlation functions. The *reduced void probability function* scales the VPF by the average number of galaxies in cells of a given size:

$$\chi(r) = \frac{-\ln(P_0(r))}{\bar{N}(r)}.$$
(2.21)

Measurements of $\bar{\xi}$, χ, and $\bar{N}$ can be used to determine which hierarchical scaling model best describes how the clustering of galaxies occurs. Different hierarchical scaling models provide different solutions to the cosmological many-body problem with varying physical justifications, and determining which best describe galaxy clustering can probe the fundamental physics behind the formation and evolution of large scale structure (eg. Saslaw and Fang 1996). In Appendix A we summarize all the

models against which we test our simulated LAE catalogs. We focus on the negative binomial model when further testing how the VPF transforms back and forth from count-in-cells.

2.4.3 The Negative Binomial Model for Hierarchical Scaling

The negative binomial model (NBM) for the hierarchical equations that govern gravitational clustering has been found to be the best fitting model for $z < 1$ galaxy samples with complete spectroscopic redshift coverage (C05; Cr04; Maurogordato and Lachieze-Rey 1987; Gaztanaga and Yokoyama 1993; Tinker $et\ al.$ 2008; Bel $et\ al.$ 2016; Yang and Saslaw 2011; Hurtado-Gil $et\ al.$ 2017; among several others) and in simulations (Andrew $et\ al.$ 2013). The negative binomial model predicts a reduced VPF (χ) from the volume-averaged two-point correlation function ($\bar{\xi}_2$) and the average number of galaxies in a cell ($\bar{N}$):

$$\chi_{\text{neg. bin.}} = \frac{\ln(1 + \bar{N}\bar{\xi}_2)}{\bar{N}\bar{\xi}_2} \tag{2.22}$$

The NBM can be derived through many methods and has been used to statistically describe phenomena across many fields of science. It is also known as the modified Bose-Einstein distribution (see Fry and Colombi 2013 for a succinct derivation in that context). The NBM was first used a cosmological context by Carruthers and Duong-van (1983) and derived for clustering analysis by Elizalde and Gaztanaga (1992). In this reworking, the probability of a galaxy appearing in a given cell depends on the number of galaxies that already exist within it, and is correlated to a uniform Poisson distribution. Or as Cr04 summarizes it, the NBM describes the probability of having a given number of 'successes' (finding a galaxy) after a certain number of 'failures' (voids), and probability of a failure (P_0) depends on the density of a sample and how clustered it is ($\bar{N}\bar{\xi}_2$). The NBM was also derived with thermodynamic arguments

by Sheth (1995) in the framework of Saslaw and Hamilton (1984), and shown to be a special case of the hyper geometric model of Mekjian (2007). Most recently, Gaztanaga and Yokoyama (1993) rederived the model by considering a sample divided into equal and independent cells and tying the cells' occupation probability to $\bar{\xi}$.

Although it is an robust description of many galaxy samples, the NBM is arguably not physically motivated. Fry and Colombi (2013) argue that "there is no fundamental reason that galaxies follow the negative binomial scaling curve, but that this follows from typical galaxy parameters" like bias and number density. Some authors find it justified (Carruthers and Duong-van 1983; Elizalde and Gaztanaga 1992; Betancort-Rijo 2000). Saslaw and Fang (1996) argued the NBM violates the second law of thermodynamics while still being the best fit to their data. Yang and Saslaw (2011) confirmed that their SDSS sample was consistent with both the NBM and quasi-equilibrium model, and preferred the quasi-equilibrium model for its physical explanation. Additionally, Yang and Saslaw (2011) found that the large cosmic variance and underestimated error from jackknife errors could explain why the negative binomial function best fit a very similar sample in C05. Later, Hurtado-Gil *et al.* (2017) confirmed that the NBM outperformed the quasi-equilibrium and other models in a blind fit of the SDSS galaxies after careful consideration of incompleteness and noise in the sample.

2.5   Fitting and Testing the NBM for LAEs

2.5.1   Fitting Count-In-Cells to the NBM

Traditionally, the reduced VPF (χ) and volume-averaged correlation functions ($\bar{\xi}_p$) are measured using the count-in-cells method. For CiC, one drops many random test spheres of a given size and counts the number of galaxies within each. The VPF

Figure 2.4: The reduced VPF, $\chi = -\log(P_0)/\bar{N}$, as a function of $\bar{N}\bar{\xi}$ for several hierarchical scaling models and our catalogs of simulated LAEs. For errors, we assume that $\bar{N}$ has no uncertainty, and we transfer the errors on the 3D VPF and $\xi_{\text{L-S}}$ to $\log(\chi)$ and $\log(\bar{N}\bar{\xi})$ respectively. The y-errors are the 1σ VPF subsample error (the shaded regions in the right side of Figure 1), and the x-errors are 3 times the $\Delta\xi = (1 + \xi_{\text{L-S}})/\sqrt{DD}$ errors (the error bars in Figure 2) scaled to $\log_{10}(\bar{\xi})$. For a completely unclustered Poisson sample, $\chi = 1$ everywhere.

can be measured with CiC by isolating the empty test spheres. The volume-averaged two-point correlation function, $\bar{\xi}_2$, is the reduced second moment of the probability distribution of the galaxies' CiC, or the variance in the number of galaxies across all test spheres i of a given size:

$$\bar{\xi}_2 = \frac{\overline{(N_i - \bar{N})^2} - \bar{N}}{\bar{N}^2} \tag{2.23}$$

As with $\bar{\xi}_2$, the higher order terms $\bar{\xi}_p$ are the reduced p'th order moments, and CiC is often used to measure the scaling coefficients S_p between the correlation functions (Croton *et al.* 2004a; Wolk *et al.* 2013; among many others). CiC has also been used to help constrain cosmological models in simulations and observations (Wang *et al.* 2019; Uhlemann *et al.* 2020; Wen *et al.* 2020; Repp and Szapudi 2020.) CiC is a versatile probe for the underlying causes of the hierarchical features that we see in large scale structure, and is also often used in other broader applications of clustering (eg. Adelberger *et al.* 1998; Mesinger and Furlanetto 2008; Jensen *et al.* 2014). As we continue this analysis, we consider how we can leverage what we have learned about the VPF's uncertainty and reliability when using CiC.

To calculate our CiC, we drop at least 500,000 test spheres at every radius, utilizing the same random placement as we used for the simple VPF and the Landy and Szalay (1993) ξ measurements. We confirm that the CiC algorithm gives identical VPF values to our previous measurement when counting empty cells, and therefore should give accurate $\bar{N}$ and $\bar{\xi}$ measurements as well. Additionally, we calculate the value of $\bar{\xi}$ at the largest radii we measured and found this integral constraint negligibly small.

The more traditional application of hierarchical scaling models has been to use CiC to plot $\bar{N}\bar{\xi}$ against the reduced VPF χ and determine what model of hierarchical scaling best fits the galaxy samples, as in C05 and Cr04. In Figure 2.4, we plot $\log_{10}(\bar{N}\bar{\xi})$ vs. χ for our catalogs and the predicted relationships from several popular

models, compiled in Appendix A from Fry and Colombi 2013, C05, and Cr04. We assume zero uncertainty in $\bar{N}$, translate the 1σ subsample-method VPF errors to $\chi(r)$, and translate the 3σ $\Delta\xi_{\text{L-S}}$ errors to $\log_{10}(\bar{N}\bar{\xi})$. By transferring our subsample method error estimation for the VPF into the reduced VPF, we can better inform the uncertainty of CiC, which might be underestimated by jackknife sampling (Yang and Saslaw, 2011).

In agreement with the work in Cr04, C05, Andrew *et al.* 2013, Hurtado-Gil *et al.* 2017, and many others, we find the NBM is the best fit model of hierarchical scaling for the $z = 3.1, 4.5, 5.7$ simulated LAEs. These catalogs' 1σ errors on χ, though, would not discount the geometric model or the gravitational quasi-equilibrium model (which is preferred over the NBM for its physical motivation in Yang and Saslaw 2011). The $z = 6.6$ catalog might be more consistent with the thermodynamic model, which treats galaxy clustering by analogy to statistical mechanics (Saslaw and Hamilton 1984; Fry 1986), but its large 1σ VPF errors do not discount the negative binomial or gravitational quasi-equilibrium models.

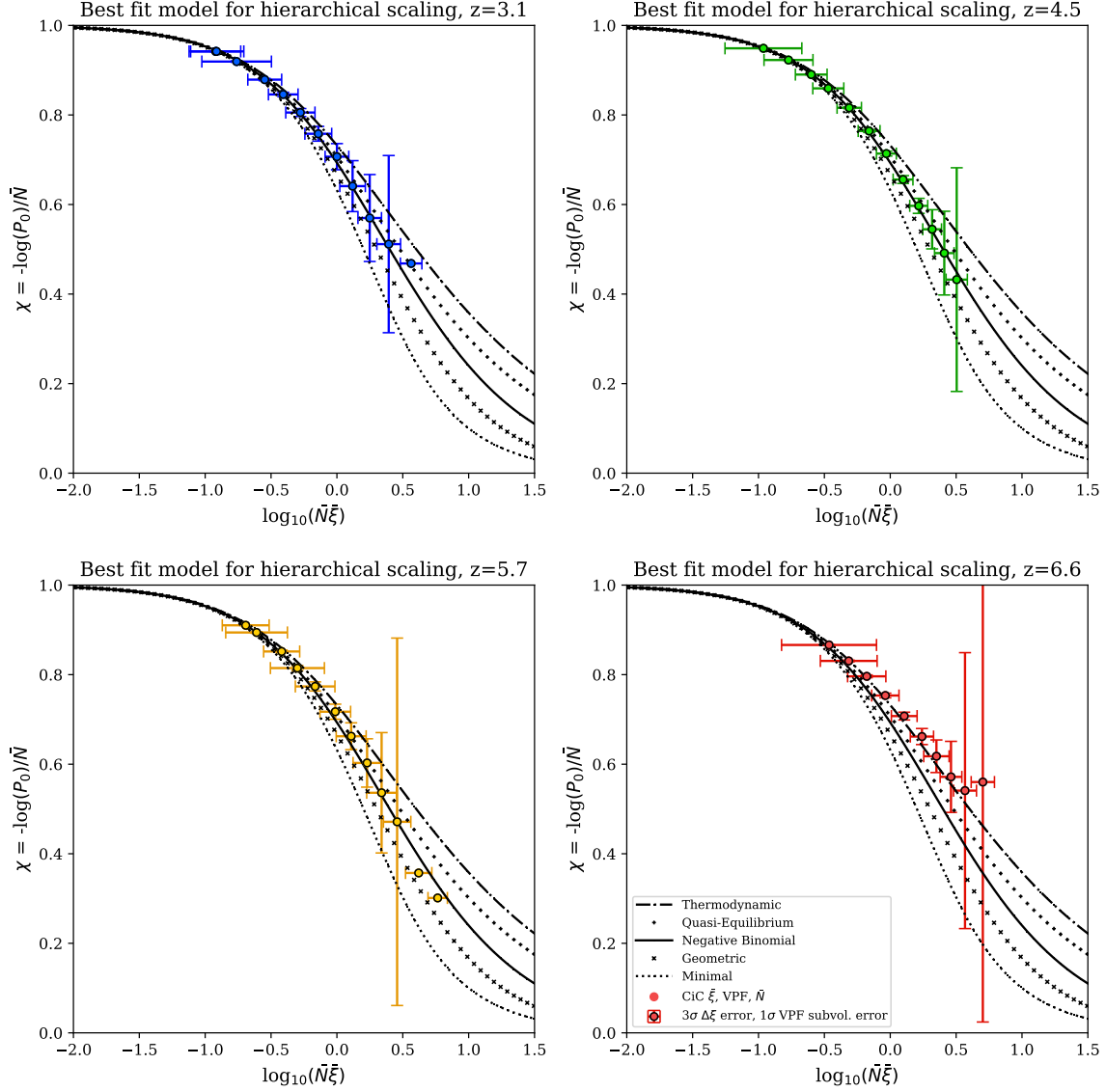**Complete Figure 2.4 caption:** The reduced VPF, $\chi = -\log(P_0)/\bar{N}$, as a function of $\bar{N}\bar{\xi}$ for several hierarchical scaling models and our catalogs of simulated LAEs. The negative binomial model is the best fit for the $z = 3.1, 4.5, 5.7$ catalogs, as expected, though the quasi-equilibrium model is not ruled out. The $z = 6.6$ catalog is perhaps most consistent with the thermodynamic model, but the large 1σ errors from the VPF subsample analysis for this smaller sample do not discount the negative binomial model or quasi-equilibrium model as the best fit. For errors, we assume that $\bar{N}$ has no uncertainty, and we transfer the errors on the 3D VPF and $\xi_{\text{L-S}}$ to $\log(\chi)$ and $\log(\bar{N}\bar{\xi})$ respectively. The y-errors are the 1σ VPF subsample error (the shaded regions in the right side of Figure 1), and the x-errors are 3 times the $\Delta\xi = (1 + \xi_{\text{L-S}})/\sqrt{DD}$ errors (the error bars in Figure 2) scaled to $\log_{10}(\bar{\xi})$. For a

completely unclustered Poisson sample, $\chi = 1$ everywhere. The consistency of the z=3.1, 4.5, and 5.7 catalogs with the negative binomial model supports the hypothesis that LAEs also follow the negative binomial model. The deviation of the LAEs at z=6.6 from the NBM and its significance requires larger simulations for confirmation at a significant level.

2.5.2 Testing Our VPF with the NBM and $\bar{\xi}_{CiC}$

To further test how well the negative binomial model fits our simulated LAEs, we use it to move between the VPF and volume-averaged two-point correlation function of our simulated LAEs. We first start with the CiC-measured $\bar{\xi}$, use the NBM to predict a VPF, and compare to the measured VPF as a test of the model fit. We combine the definitions of the reduced VPF and the NBM in Equations 2.21 and 2.22 to yield the following form of the model, which we can solve with various root-seeking algorithms at each radius.[4] This transformation allows us to derive a volume averaged two-point correlation function from the VPF of the T09 simulated LAEs, or vice versa:

$$0 = P_0 \ (1 + \bar{N}\bar{\xi}_2)^{1/\bar{\xi}_2} - 1 \tag{2.24}$$

In their work, C05 measured the VPF, $\bar{\xi}$, and $\bar{N}$ with CiC. To further separate the VPF from the CiC and two-point correlation function, we approximate $\bar{N}(r)$ (and therefore χ) by assuming homogeneous distribution at the same number density of the catalogs:

---

[4]We use *scipy.optimize.root* and the Levenberg-Marquardt sub-module (Jones *et al.*, 01) for its stability, and verify the answers with manual evaluations. We also remove abnormally large values generated at numerically unstable regions, and limit ourselves to $|\bar{\xi}_{\text{VPF}}| < 3$.
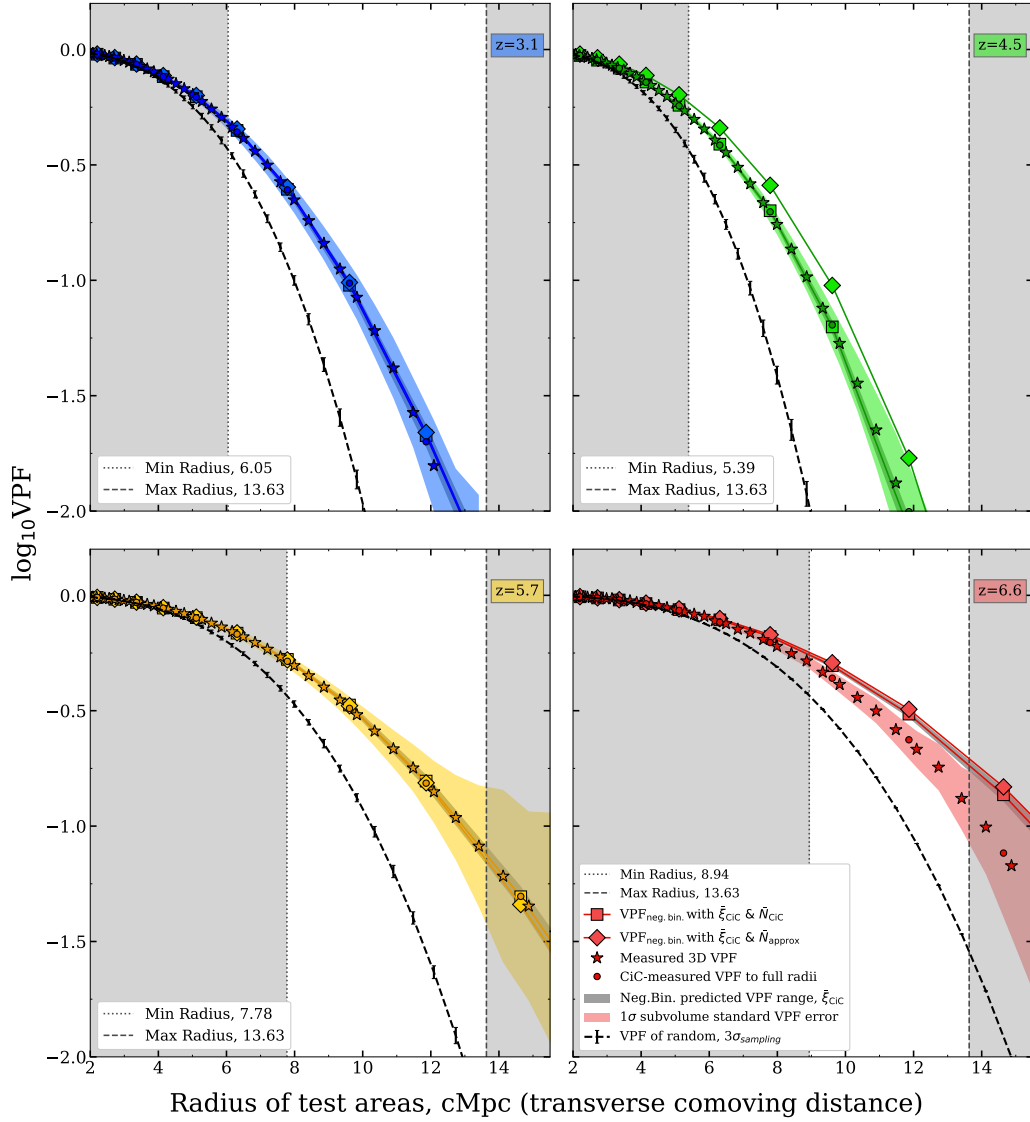
Figure 2.5: Comparing the measured 3D VPF Vs. the VPF derived from the NBM using the measured Count-in-Cells. The dark colored stars are the measured 3D VPF, and the shaded colored regions indicate the 1σ standard error across the independent $(51\text{ cMpc})^3$ sub-volumes' VPFs. The dark colored circles are the VPF measured with the CiC. The colored squares (diamonds) are the derived VPF using $\bar{\xi}_{\text{CiC}}$ and $\bar{N}_{\text{CiC}}$ ($\bar{N}_{\text{approx}}$). The grey shading around the $\bar{\xi}_{\text{CiC}}$ & $\bar{N}_{\text{CiC}}$ squares comes from carrying $\pm 1\sigma = \Delta\xi = (1 + \xi_{\text{L-S}})/\sqrt{DD}$ errors through the negative binomial model.

46

$$\bar{N}_{\text{approx}}(r) \approx \mathcal{N} \, \frac{4}{3}\pi R^3 \quad \rightarrow \quad \chi(r) \approx \frac{-\ln(P_0(r))}{\mathcal{N}\,(4/3)\pi R^3}. \tag{2.25}$$

We note that our shortcut to the reduced VPF is reminiscent of the normalized VPF value explored by Maurogordato and Lachieze-Rey (1987), in which n is the number volume density and V is any given volume:

$$\chi_{\text{M-L,1987}} = \frac{\ln(P_0)}{nV} \tag{2.26}$$

We find that this $\bar{N}_{\text{approx}}$ is significantly similar to the true measured value of $\bar{N}_{\text{CiC}}$, and our final results are not affected by this choice. This similarity is to be expected, as both attempt to measure an average of galaxies in a given volume, though $\bar{N}_{\text{approx}}$ and $\bar{N}_{\text{CiC}}$ might diverge in situations where the shape of the one-point count-in-cells distribution is very skewed (as in some of the cosmological Quijote simulations tested in Uhlemann *et al.* 2020 using the probability density function of the entire matter field). For us, the only notable difference between $\bar{N}_{\text{approx}}$ and $\bar{N}_{\text{CiC}}$ is in the $z = 4.5$ catalog, where using $\bar{N}_{\text{approx}}$ raises $\log_{10}(\bar{\xi}_{\text{VPF}})$ by about the width of the 1σ VPF errors in Figure 2.5.

First, we follow the typical application of the hierarchical scaling models to derive a VPF from the CiC measurements and then compare to the measured VPF in Figure 2.5. We assume a NBM to transform the volume averaged two-point correlation function into a VPF, the more traditional use of hierarchical scaling as seen in C05 and Cr04. We compare how well the volume averaged form from CiC ($\bar{\xi}_{\text{CiC}}$) recreates the VPF under the NBM, with either the true measured average number of galaxies from CiC ($\bar{N}_{\text{CiC}}$) or an approximation from number density ($\bar{N}_{\text{approx}}$). As Figure 2.4 predicted, using all CiC measurements for $\bar{\xi}$ and $\bar{N}$ reproduces the VPF exactly for the $z = 3.1, 4.5, 5.7$ catalogs. We verify that the shortcut of using $\bar{N}_{\text{approx}}$ over $\bar{N}_{\text{CiC}}$

does not influence the final conclusions. The deviation of the CiC values from the NBM in $z = 6.6$ lie at the edge large 1σ errors about the VPF measurement, nearly overlapping with the transformed 1σ $\Delta\xi_{\text{L-S}}$ errors about the NBM-predicted VPF.

Complete Figure 2.5 caption: Comparing the measured 3D VPF vs. the VPF derived from the negative binomial model using the measured count-in-cells volume averaged two-point correlation function. The dark colored stars are the measured 3D VPF of each volume, and the shaded colored regions indicate the 1σ standard error across the independent $(51 \text{ cMpc})^3$ sub-volumes' VPFs. The dark colored circles are the VPF measured with the CiC, and the excellent match with our independent measurement confirms our algorithm is correct. The colored squares (diamonds) are the derived VPF using $\bar{\xi}_{\text{CiC}}$ and $\bar{N}_{\text{CiC}}$ ($\bar{N}_{\text{approx}}$). Our final results do not change with our choice for $z = 3.1, 5.7, 6.6$, and very slightly change for $z = 4.5$, indicating that approximating $\bar{N}$ from the number density is is a valid shortcut to measuring it with CiC. The grey shading around the $\bar{\xi}_{\text{CiC}}$ & $\bar{N}_{\text{CiC}}$ squares comes from carrying $\pm 1\sigma = \Delta\xi = (1 + \xi)/\sqrt{DD}$ errors from $\xi_{\text{L-S}}$ through the negative binomial model. The black dashed lines indicate the predicted VPF for the simulations were they completely randomly distributed. The maximum and minimum radii are theoretically derived from the simulations' number densities, volumes, and our chosen VPF precision of $\pm 10^{-2}$. As Figures 2.4 and 2.6 also show, the negative binomial model transformations using the CiC agree excellently with the observed clustering signals for $z = 3.1, 4.5, 5.7$. The z=6.6 LAEs deviate from the negative binomial model predictions, lying just past the 1σ error bounds of the measurements.

2.5.3   Deriving $\bar{\xi}$ from the VPF with the NBM

Past studies like those of C05, Cr04, and Andrew *et al.* (2013) measured two point and higher order correlation functions, derived the VPF assuming given models for

the hierarchical scaling, and then compared against measured VPF to compare the models. Our work here builds off their key results–the strength of the NBM to predict the VPF of low-redshift galaxies–and inverts this pattern in our simulations of high-redshift LAEs as an additional test of the NBM. We begin from a measured VPF, assume the NBM to derive a correlation function, and then compare to the measured CiC correlation function. This serves as an extra step to test the NBM, the behavior of CiC for different samples, and the shortcut of using a density-approximated $\bar{N}_{\mathrm{approx}}$ over the directly CiC-measured $\bar{N}_{\mathrm{CiC}}$. We find that deviations from NBM, especially in the $z = 6.6$ catalog, appear more obvious when moving from the VPF into $\bar{\xi}$ than in the opposite direction, thanks to the transformation into straight lines in log-log space.

We solve the NBM in Equation 2.24 for $\bar{\xi}_{\mathrm{VPF}}$ in Figure 2.6 given our measured VPF P_0 and either $\bar{N}_{\mathrm{CiC}}$ or $\bar{N}_{\mathrm{approx}}$. We compare the derived $\bar{\xi}_{\mathrm{VPF}}$ (colored stars and crosses, depending on the choice of $\bar{N}$) and the CiC-measured $\bar{\xi}_{\mathrm{CiC}}$. We show the ranges corresponding to the power law fits to the $\bar{\xi}_{\mathrm{VPF}}$ from the VPF $\pm 1\sigma$ subsample uncertainties, which straddle the best-fit power law for our central $\bar{\xi}_{\mathrm{VPF}}$ values. Figure 2.6 agrees with Figures 2.4 and 2.5: the $z = 3.1, 4.5, 5.7$ catalogs show excellent consistency with the NBM, as their $\bar{\xi}_{\mathrm{VPF}}$ are exactly consistent with $\bar{\xi}_{\mathrm{CiC}}$. The unique behavior of the $z = 6.6$ catalog might be clarified here, showing the discrepancy from the NBM in Figure 2.4 centers around the LAEs' $\bar{\xi}_{\mathrm{CiC}}$ not behaving as the NBM predicts. This might be due to the fact that the $z = 6.6$ catalog is the smallest and most clustered, perhaps indicating that higher clustering amplitudes might make our approximation for $\bar{N}$ become invalid earlier and create larger deviations from the true hierarchical scaling description. The reasons for the different behavior of the z=6.6 LAE catalog require larger samples and larger simulations to confirm and understand. We leave that for future work.
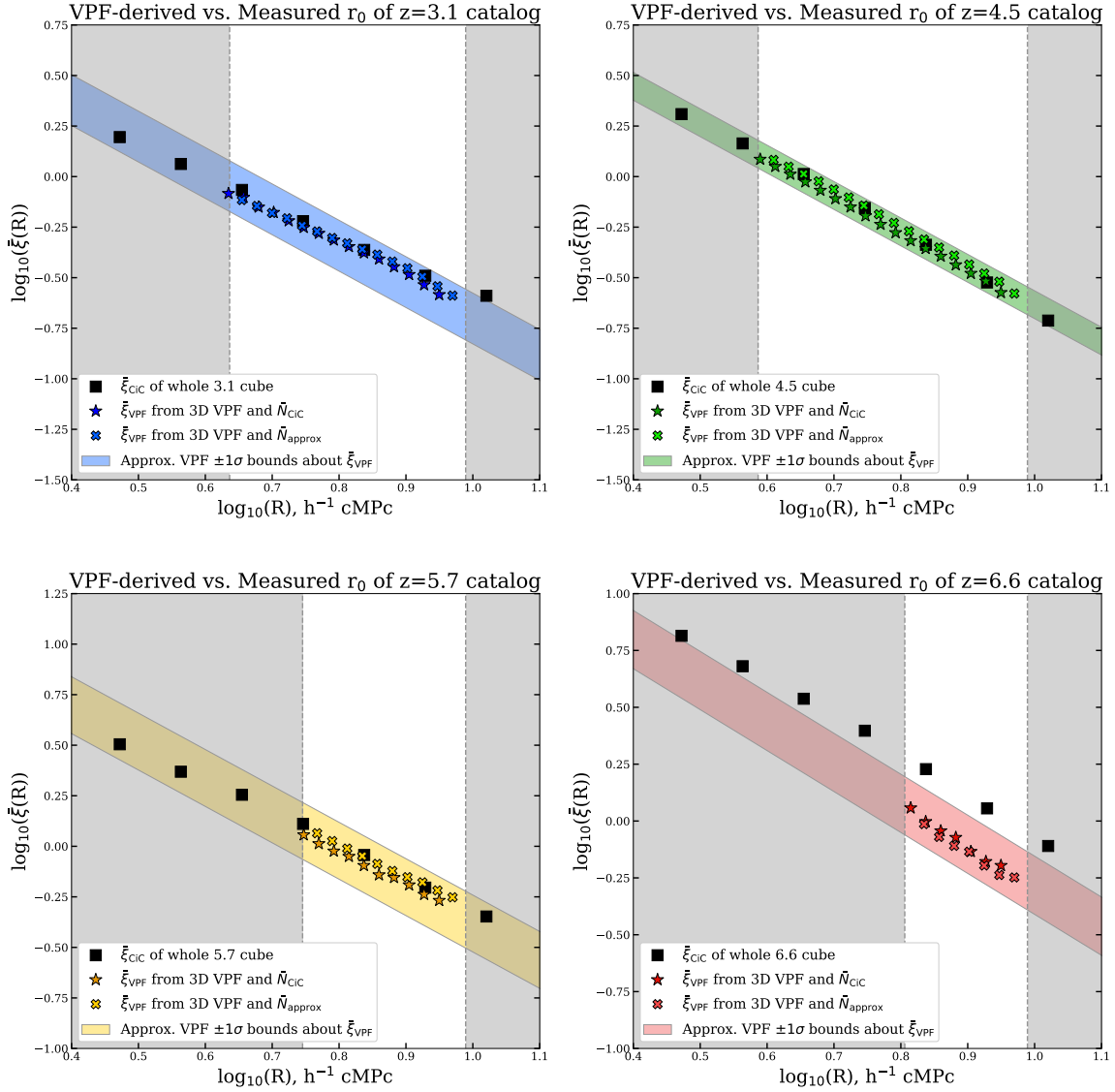
Figure 2.6: We assume a negative binomial model to transform the VPF into a volume-averaged two-point correlation function, and compare $\bar{\xi}_{\mathrm{VPF}}$ to the directly measured $\bar{\xi}_{\mathrm{CiC}}$. The black squares are $\bar{\xi}_{\mathrm{CiC}}$ calculated using Equation 2.23. The colored stars (crosses) are the derived $\bar{\xi}_{\mathrm{VPF}}$ from the central 3D VPF of the simulations, using the CiC-measured (density-approximated) $\bar{N}$. The colored shaded regions correspond the power law fits to the $\bar{\xi}_{\mathrm{VPF}}$ from the VPF $\pm 1\sigma$ subsample uncertainties.

We again confirm that three of the four catalogs show great agreement with the NBM and the last good agreement with the quasi-equilibrium model (the next best-fitting hierarchical scaling model), agreeing with many low-redshift galaxy observations. Therefore, we show that the approximating the average number of galaxies in a cell $\bar{N}$ using number density and assuming the NBM allows us to accurately recreate the CiC volume-averaged two-point correlation function only using the VPF. Additionally, our $\bar{\xi}_{\mathrm{VPF}}$ carry through the intuitive errors that we derived in §2.3.3, and avoid the possibly underestimated errors often given to CiC measurements (Yang and Saslaw, 2011).

Complete Figure 2.6 caption: Comparing the directly measured $\bar{\xi}_{\mathrm{CiC}}$ vs. the $\bar{\xi}_{\mathrm{VPF}}$ from transforming the VPF using the negative binomial model. The black squares are $\bar{\xi}_{\mathrm{CiC}}$ calculated using Equation 2.23. The colored stars (crosses) are the derived $\bar{\xi}_{\mathrm{VPF}}$ from the central 3D VPF of the simulations, using the CiC-measured (density-approximated) $\bar{N}$. The colored shaded regions correspond the power law fits to the $\bar{\xi}_{\mathrm{VPF}}$ from the VPF $\pm 1\sigma$ subsample uncertainties, and straddle the best-fit power law for our central $\bar{\xi}_{\mathrm{VPF}}$ values. We calculate $\bar{\xi}_{\mathrm{VPF}}$ within the theoretical radii derived and discussed in §2.3.3 (grey regions), based on the catalogs' number densities, volumes, and our chosen precision of $\pm 10^{-2}$. The pattern of clustering between the catalogs translates through $\bar{\xi}_{\mathrm{VPF}}$. — The $\bar{\xi}_{\mathrm{VPF}}$ matches the $\bar{\xi}_{\mathrm{CiC}}$ very well for the $z = 3.1, 4.5, 5.7$ catalogs, confirming the negative binomial is the best fitting scaling model for those catalogs. The $z = 6.6$ catalog $\bar{\xi}_{\mathrm{VPF}}$ and $\bar{\xi}_{\mathrm{CiC}}$ appear to disagree here (perhaps due to this catalog's size and strong clustering), though Figure 2.4 and the 1σ VPF bounds suggest this deviation from the negative binomial model might not be statistically significant. The choice of $\bar{N}$ makes little difference to measuring $\bar{\xi}$ with $\bar{\xi}_{\mathrm{VPF}}$, meaning that the number density approximation with the VPF can serve as a shortcut to measuring CiC for large samples.

2.6 Conclusion

The Void Probability Function is an underutilized measure of clustering that inherently involves the higher order correlation functions (White 1979; Maurogordato and Lachieze-Rey 1987). It contains additional information about large scale structure and higher order clustering that the two-point correlation function cannot discern. When it has been used, the VPF has been able to fit low-redshift galaxy observations to a negative binomial model for the hierarchical gravitational clustering moments (eg. Croton *et al.* 2004a; Conroy *et al.* 2005). For many years, the VPF was treated as a uniquely capable measurement of clustering for very small galaxy samples (Palunas *et al.* 2004; Kashikawa *et al.* 2006), and as a powerful probe for dark matter halo occupation modeling (Fry 1986; Maurogordato and Lachieze-Rey 1987; Little and Weinberg 1994; Berlind and Weinberg 2002). In this work, we attempt to determine and test the limits of what the VPF is capable, utilize it to probe higher-order clustering in LAEs, and offer ways it can inform clustering analyses with count-in-cells and the two-point correlation function.

We measure the 2D and 3D void probability function of four simulated catalogs of Lyman-α emitters at redshifts 3.1, 4.5, 5.7, and 6.6 from the work of Tilvi *et al.* (2009). We offer general guidelines for understanding and quantifying the uncertainty in the VPF. We present suggested limits to when the VPF can be effectively studied, depending on the number density of the sample, the volume of the survey, and the minimum VPF value to be studied. We choose the independent subsample method to approximate the error in the VPF after finding that jackknife sampling understates the uncertainty of the VPF. We divide each volume into the eight independent cubes of $(51 \text{ cMpc})^3$ and measure the VPF on each sub-volume and the standard error across them as a 1σ error on the VPF. These simulated LAE catalogs and our

VPF measurements can serve as external 'correct' error estimators for observations of LAEs.

Next, we measure the standard Landy and Szalay (1993) and volume averaged Count-in-Cells two-point correlation functions for the catalogs. We re-affirm the $z = 3.1, 5.7, 6.6$ and improve the $z = 4.5$ correlation lengths measured in T09, and explore the transformation between ξ and $\bar{\xi}$ for our catalogs. Then, we compare how different models of hierarchical scaling between the volume-averaged correlation functions compare to the catalogs' measured $\bar{N}_{\text{CiC}}$, $\bar{\xi}_{\text{CiC}}$, and VPF. We find the $z = 3.1, 4.5, 5.7$ catalogs are best fit by the negative binomial model, though the uncertainties on the VPF do not rule out the quasi-equilibrium model. We find that the $z = 6.6$ might be better fit by the thermodynamic or quasi-equilibrium models, though the uncertainties on our VPF values do not discount the negative binomial model.

We further explore how the negative binomial model fits our simulated LAEs by transforming $\bar{\xi}_{\text{CiC}}$ into a VPF, and then the VPF into a $\bar{\xi}_{\text{VPF}}$. We confirm that approximating $\bar{N}$ with the number density rather than using the measured $\bar{N}_{\text{CiC}}$ does not change our final results. We find all four catalogs' $\bar{\xi}_{\text{CiC}}$ predict accurate VPF values with the negative binomial model (within both the 1σ VPF and 1σ $\Delta\xi$ errors). We find that the $z = 3.1, 4.5, 5.7$ catalogs' VPFs derive very accurate $\bar{\xi}_{\text{VPF}}$ with the NBM, while the $z = 6.6$ VPF yields a lower-than-expected $\bar{\xi}_{\text{VPF}}$ (though still within our uncertainties). The deviation of the NBM at our highest redshift might perhaps explained by the fact that $z = 6.6$ is our sparsest and most clustered catalog. This behavior requires larger simulations to fully understand. These results indicate that high-redshift starburst galaxies like LAEs in the T09 model show large-scale clustering behavior similar to that of local galaxies, also following the negative binomial model. This suggests the LAEs likely do not have excessive clustering in higher-order

correlation terms, though the size of our simulations do not let us decisively rule out the presence of higher-order clustering nor identify additional redshift dependence.

Our guidelines of when the VPF is best used offer greater use beyond improving future clustering analyses in moderately sized samples. In this age of growing astronomical data, there are many problems that would respond positively to fast and intuitive tools with well-explored uncertainties. We have shown, as expected, that the VPF conserves the same core qualitative conclusion of the standard Landy and Szalay (1993) two-point correlation function through the NBM transformation: the clustering of LAEs increases with redshift. This overlap allows us to consider which clustering applications might prefer the brevity of calculating the VPF and its inherent connection to higher-order clustering. For example, the upcoming Nancy Grace Roman Space Telescope (previously known as WFIRST) will observe many thousands of LAEs during the Epoch of Reionization ($z > 6$). Due to the resonant properties of the Lyα line in neutral hydrogen, the prevalence and clustering of LAEs is able to track the amount of neutral hydrogen and its distribution around them (eg. McQuinn *et al.* 2007; Jensen *et al.* 2013; Zheng *et al.* 2016; Kakiichi *et al.* 2016; Ouchi *et al.* 2018). Samples in more neutral IGM will be more clustered, both because their emission is more attenuated at higher neutral fractions, but also because LAEs are more likely to be observed in regions that have already been ionized (eg. Furlanetto *et al.* 2006). For this application, the VPF might be a useful tool (Gangolli *et al.* 2021, Perez et al. in prep)–it easily compares clustering between samples of the same density, is more efficient to calculate than CiC, and has intuitive limits that can help guide the planning of observations.

Chapter 3

PROBING PATCHY REIONIZATION WITH THE VOID PROBABILITY FUNCTION OF LYMAN-ALPHA EMITTERS

Lucia A. Perez[1], Sangeeta Malhotra[2], James E. Rhoads[2], and Peter Laursen[3]

[1] Arizona State University School of Earth and Space Exploration, Tempe, AZ

[2]NASA Goddard Space Flight Center, Greenbelt, MD,

[3]Cosmic Dawn Center (DAWN) and Niels Bohr Institute, University of Copenhagen

Jagtvej 128, 2200 Copenhagen N, Denmark

A paper based on this chapter has been submitted to the Astrophysical Journal; please see the refereed publications for updated conclusions based on peer review.

ABSTRACT

We probe what constraints for the global ionized hydrogen fraction the Void Probability Function (VPF) clustering can give for the Lyman-Alpha Galaxies in the Epoch of Reionization (LAGER) narrowband survey as a function of area. Neutral hydrogen acts like a fog for Lyman-alpha emission, and measuring the drop in the luminosity function of Lyman-α emitters (LAEs) has been used to constrain the ionization fraction in narrowband surveys. However, the clustering of LAEs is independent from the luminosity function's inherent evolution, and can offer additional constraints for reionization under different models. The VPF measures how likely a given circle is to be empty. It is a volume-averaged clustering statistic that traces the behavior of higher order correlations, and its simplicity offers helpful frameworks for planning surveys.

Using the Jensen *et al.* (2014) simulations of LAEs within various amount of ionized intergalactic medium, we predict the behavior of the VPF in one ($301 \times 150.5 \times 30$ Mpc^3), four (5.44×10^6 Mpc^3), or eight (1.1×10^7 Mpc^3) fields of LAGER imaging. We examine the VPF at 5 and 13 arcminutes, corresponding to the minimum scale implied by the LAE density and the separation of the 2D VPF from random, and the maximum scale from the 8-field 15.5 deg^2 LAGER area. We find that even a single DECam field of LAGER (2-3 deg^2) could discriminate between mostly neutral vs. ionized. Additionally, we find four fields allows the distinction between 30, 50, and 95 percent ionized; and that eight fields could even distinguish between 30, 50, 73, and 95 percent ionized.

3.1   Introduction

Lyman-α emitters (LAEs) are unique galaxies whose properties make them excellent probes of various phenomena at high redshift. High-redshift LAEs can be selected using narrowband imaging in the optical and infrared, and are particularly suited to study the epoch of reionization (EoR). The EoR was a key era and final phase change of the universe where the opaque 'fog' of early neutral hydrogen was ionized by the earliest galaxies. By leveraging the strong emission in the resonant Ly-α line, the observed properties and distribution of $z > 6$ LAEs can be used to understand the extent of neutral hydrogen around them and the process and pacing of reionization (Miralda-Escudé 1998; Rhoads and Malhotra 2001; Hu *et al.* 2002; Furlanetto *et al.* 2006; Mesinger and Furlanetto 2008; McQuinn *et al.* 2007).

LAEs sensitively inscribe the presence of neutral hydrogen around them as it attenuates their brightness and luminosity function (e.g. Haiman and Spaans 1999; Santos 2004; Furlanetto *et al.* 2004), and the extent of ionized hydrogen as it reveals clusters of LAEs (e.g. Hu *et al.* 2021; Sobral *et al.* 2015; and others). Much work has been done to constrain and understand reionization using the evolution of the Lyα luminosity function (LF) across redshift history (e.g. Ajiki *et al.* 2003; Ouchi *et al.* 2003; Hu *et al.* 2004; Malhotra and Rhoads 2004b; Ouchi *et al.* 2010; Santos *et al.* 2016; Ouchi *et al.* 2018; Morales *et al.* 2021). McQuinn *et al.* (2007) modeled the expected suppression of the Lyα LF across many neutral fractions for the global IGM, contributing to constraints of neutral hydrogen fraction to between 0.2-0.4 at $z \sim 7$ (seen in other LAE surveys, e.g. Konno *et al.* 2018). Similar modeling from Santos (2004) contributed to similar constraints in Ouchi *et al.* (2010). Some works have found no evidence of attenuation in the $z > 6$ Lyα LF in their analysis (Malhotra and Rhoads 2004b; Tilvi *et al.* 2010); while some analyses have measured

an attenuation in the LFs or LAE number densities within $5.7 < z < 7$ (Kashikawa *et al.* 2006; Iye *et al.* 2006; Ouchi *et al.* 2010; Kashikawa *et al.* 2011; Konno *et al.* 2014, 2018).

The Lyman-Alpha Galaxies in the Epoch of Reionization (LAGER) survey is the largest narrowband survey yet of LAEs during reionization, on track to cover $\sim$24 deg^2 and better constrain the timing and morphology of reionization. The survey utilizes a unique/specially made/proprietary N964 narrowband filter centered at 9642Å with a filter FWHM of 92Å (Zheng *et al.*, 2019) that exploits the 3 deg^2 field of view of the Dark Energy Camera (DECam) instrument mounted on the Blanco 4m telescope at Cerro Tololo International Observatory. LAGER has to date observed 195 z=6.9 LAEs across four fields and 10.19 deg^2 (Zheng *et al.* 2017; Hu *et al.* 2019; Wold *et al.* 2021), making it the largest survey of LAEs near z$\sim$7 to date with four additional fields still to be analyzed or imaged. With a high rate of spectroscopic confirmation of LAEs (Hu *et al.* 2017; Yang *et al.* 2019; Harish et al. submitted) and a survey volume at least twice as large as others used in Lyα LF analyses (Wold *et al.*, 2021), LAGER is a powerful and efficient survey of Lyα during reionization. So far, the LAGER Lyα luminosity functions and their evolution are consistent with a nearly completely ionized universe at z=6.9 (Hu *et al.* 2019; Wold *et al.* 2021).

Neutral fraction constraints from studies of the Lyα LF can show tension with the measurements of the IGM temperature (which indicate a mostly ionized medium at $z \sim 6$, e.g. Fan *et al.* 2006) and measurements of the Lyα line profile (e.g. Ouchi *et al.* 2010 rule out a fully neutral universe at $z = 6.6$ with the model of Haiman and Cen 2005). However, the Lyα LF analyses come with the caveat that the LF suppression might be explained by the evolution of the halo mass function (Dijkstra *et al.*, 2007), cosmic variance, details of the chosen model, or changes in when galaxies show strong Lyα emission (e.g. Ota *et al.* 2008; Stark *et al.* 2010; Pentericci *et al.* 2011; Ono *et al.*

2012; Schenker *et al.* 2012; Endsley *et al.* 2021). A way to break these tensions and degeneracy is with the clustering of LAEs during the epoch of reionization, which offers an additional way to constrain reionization that can circumvent possible galaxy evolution (McQuinn *et al.*, 2007) and is independent from the evolution of the intrinsic LF.

In this work, we focus on what constraints for reionization the Void Probability Function (VPF) might give for the growing collection of $z = 6.9$ narrowband selected LAEs from LAGER. Specifically, we seek to quantify what precision on the constraints of reionization the LAGER survey may find with clustering, as a function of imaged area. The VPF is a statistical measure of clustering that simply asks: how likely is a circle or sphere of a given size to be empty in the sample? It is also the zero-point volume-averaged correlation function, and is measured by counting how many cells of a given size are empty, often alongside count-in-cells. As the 0th moment of count-in-cells, it carries the signature of higher-order correlation functions beyond two-points, and its simplicity can be leveraged to guide the number density and volume of surveys (Perez *et al.*, 2021). Kashikawa *et al.* (2006), McQuinn *et al.* (2007), and Gangolli *et al.* (2021) have examined the ability of the VPF to constrain reionization for various generations of Subaru LAE observations at $z = 6.6$ and $z = 5.7$. Here, we focus on the constraints and survey guidelines that the VPF can give for the uniquely large and growing LAGER survey of LAEs at $z = 6.9$ in tandem with the LAGER luminosity function analysis.

This paper is organized as follows. In §3.2, we describe our use of the Jensen *et al.* (2014) simulations of LAEs within different IGM fractions of neutral hydrogen to predict the VPF of LAGER LAEs. In §3.3, we introduce the use of clustering to constrain the fraction of neutral hydrogen during the epoch of reionization, and motivate the focus on the Void Probability Function for the LAGER narrowband survey.

59

In §3.4, we ask and answer: how distinguishable are the different ionization fractions using the VPF for narrowband-detected LAGER LAEs, as a function of survey area? In particular, we leverage the large simulation volume at various ionization fractions to measure the VPF in mock LAE slices that mimic a single LAGER field, the four currently imaged, and the eight in the full survey plan. Our work and conclusions are summarized in §3.5.

3.2   Simulations and Methodology

3.2.1   The Jensen et al. (2014) Simulations of LAEs During Reionization

To obtain the ionization fraction and the Lyα luminosities of galaxies, we make use of the simulations of Jensen *et al.* 2014 who modeled LAEs during the epoch of reionization (expanding upon Jensen *et al.* 2013). The simulations exist at the specific mass-averaged ionized hydrogen fractions in the IGM of $\langle x_i \rangle_m = \{0.30, 0.50, 0.58, 0.73,$ $0.83, 0.92, 0.95\}$, and for our applications, give each LAE's transmitted Lyα luminosity. The strength of this model is its combination of a large volume, ensuring a statistically sound sample of galaxies, and high-resolution radiative transfer. For example, we find the cosmic variance across the entire Jensen *et al.* (2014) volumes is less than 4% after all selections using the Trenti and Stiavelli (2008) cosmic variance calculator [1]. Here we briefly summarize the model, and direct readers to those works for full details.

The large-scale structure of the Universe was modeled using a 165 billion particle N-body simulation of a (602x607x600) Mpc^3 volume. Halos were populated with galaxies to match the UV Lyman-break and Lyα luminosity functions of Ouchi *et al.* (2010). Each galaxy was modelled as emitting an intrinsic, halo mass-dependent,

---

[1] *https://www.ph.unimelb.edu.au/~mtrenti/cvc/CosmicVariance.html*

double-peaked Lyα spectrum. The intrinsic Lyα luminosity that emerges from a given LAE at $1.5r_{vir}$ was randomly drawn from a lognormal distribution with a standard deviation of $\sigma = 0.4$ dex and a mean proportional to the host dark matter halo's mass. Details of the unique "Gaussian-minus-Gaussian" line shape recipe used to generate the double-peaked spectra can be found in Jensen *et al.* (2013). Subsequently, the individual Lyα spectra were modified to account for scattering in the IGM, using the radiative transfer (RT) code IGMTRANSFER (Laursen, 2011). The RT code calculates the transmission across the Lyα line through the circum- and intergalactic medium, on the basis of the much higher-resolution, hydrodynamic, cosmological simulations of Laursen *et al.* (2011). The wavelength-dependent transmission is defined as the median value of sightlines in all directions, of all galaxies in the simulation. Sightlines began at a distance of 1.5 virial radii from the center of a given galaxy, at which distance most of the Lyα photons have experienced their last scattering *into* the line of sight, after which they are only scattered *out of* the line of sight (Laursen *et al.*, 2011). The resulting direction-dependent transmitted Lyα luminosities are the selection criteria for our LAEs.

Next, we highlight these and other assumptions that Jensen *et al.* (2014) made in their single simulated reionization history. When simulating the radiative transfer of ionizing radiation out of galaxies, each source from the N-body simulation was given a flux proportional to its mass. Therefore, the photoionization rate of each source (Eq. 1 in Jensen *et al.* 2013) depends on assumptions for the initial mass function, star formation efficiency, and escape fraction of the galaxies. Additionally, Jensen *et al.* (2013) assume that only galaxies took part in reionization, and that small sources with $M_h < 10^9 M_\odot$ were suppressed once the IGM around them was more than 10% ionized (a.k.a. self-regulation). This model therefore generated *inside-out reionization*, where the first regions to be ionized are those that

have the most galaxies and are the highest mass (confirmed by the higher mass-averaged ionization fractions than volume-averaged at a given redshift). The Jensen *et al.* (2013) and Jensen *et al.* (2014) simulations did not include the effects of gas clumping and Lyman limit systems, and give the disclaimer that they may have overestimated the redshift when reionization ended. They also assumed that the true topology of reionization would be consistent with their model (citing Friedrich *et al.* 2011 and Iliev *et al.* 2012), and therefore analyzed and shared their simulations with *mass-averaged* ionization fractions rather than volume-averaged. The mass-averaged ionization fractions correspond approximately to volume-averaged ionized hydrogen fractions of $\langle x_i \rangle_v = \{0.22, 0.40, 0.485, 0.66, 0.78, 0.89, 0.93\}$, according to Fig. 2 of Jensen *et al.* (2013).

Various observations were used to fine tune and calibrate the assumptions made when creating these simulations. To study the volume at a particular ionization fraction, they selected whatever redshift corresponds to the sought fraction and scale the dark matter halo masses so that the intrinsic luminosity function of the volume still matched the observed Lyα LF at z = 5.7 (Ouchi *et al.*, 2010). The simulation is by default scaled to $z = 6.5$, which allows our analysis for LAGER at $z = 6.9$. The rest equivalent width (EW) distribution followed a lognormal EW distribution supported by lower redshift observations (e.g. Reddy and Steidel 2009, as it was not well-known at z=6.5). Galaxies are randomly assigned EWs with no correlation to Lyα luminosity. The EW distribution and assignment was found to be consistent with (Stark *et al.* 2010; Jiang *et al.* 2013; Hayes *et al.* 2014; Zheng *et al.* 2014). Jensen *et al.* (2013) and Jensen *et al.* (2014) took the z=5.7 Lyα luminosity function of Ouchi *et al.* (2010) as true, and match the mass-to-light ratio of their luminosity to it.

With regards to LAE clustering, Jensen *et al.* (2013) first ran smaller simulations with this model, and compared their analyses of the simulated LAEs to: the observed

Lyα luminosity functions of Ouchi *et al.* (2010) and Kashikawa *et al.* (2011); measured decreases in the observed LAE fraction; and the 3D and 2D two-point correlation functions of observed LAEs (measured $\xi(r)$ and its projection into $w(\theta)$; Ouchi *et al.* 2010), all as a function of $\langle x_i \rangle_m$. With expanded simulations, Jensen *et al.* (2014) showed that CiC could tell apart ionization fractions and probe the unique signal of inhomogeneous reionization on the clustering of LAEs (McQuinn *et al.*, 2007). We add the VPF to the repertoire of clustering statistics used on these simulations, alongside the rest of the work we present.

Throughout this work, we use transmitted Lyα luminosity in the z-direction in any place we refer to the luminosities of the simulated LAEs. We use the *astropy*[2] cosmology package under a Planck 2013 flat ΛCDM cosmology, with $\Omega_{M,0}$=0.307, $\Omega_{b,0}$=0.0483, H_0=67.8 km (Mpc s)$^{-1}$. The original simulations were run with a flat Λ cold dark matter consistent with the 9 year Wilkinson Microwave Anisotropy Probe model with $(\Omega_m, \Omega_b, \text{h}, \text{n}, \sigma_8) = (0.27, 0.044, 0.7, 0.96, 0.8)$, WMAP results (Hinshaw *et al.*, 2013), and we note this difference in cosmology creates a negligible $> 0.5\%$ change in the surface densities and radii we list.

3.2.2 Creating Mock LAGER LAE Volumes for VPF Predictions

As the LAGER collaboration prepares for the clustering analysis of our LAEs and expands observations to more fields, we consider: what additional and complementary constraints can the VPF give for the global ionization fraction during the epoch of reionization? In this section, we parse up the (~ 600 cMpc)^3 volume of Jensen *et al.* (2014) into subslices that mimic the breadth of the LAGER at $z = 6.9$, to prepare to create predictions for the 2D VPF clustering of the narrowband-selected LAEs of

---

[2]*https://www.astropy.org*: a community-developed core Python package for astronomy (Astropy Collaboration *et al.*, 2013, 2018).

Table 3.1: Our Use of the Jensen *et al.* (2014) Simulations of LAEs in Reionization for the LAGER VPF Tests.

$\langle x_i \rangle_m$	$\langle x_i \rangle_v$	N_{total}	N_{selected}	Volume Density $\mathcal{N}_{\text{Mpc}^{-3}}$	Surface Density $\Sigma_{\text{amin}^{-2}}$
4-field	LAGER	...	195	2.86×10^{-5}	4.74×10^{-3}
0.30	0.22	978,687	2,461	1.12×10^{-5}	2.20×10^{-3}
0.50	0.40	1,632,805	3,649	1.66×10^{-5}	3.26×10^{-3}
0.73	0.66	2,301,037	6,378	2.91×10^{-5}	5.7×10^{-3}
0.95	0.93	3,052,695	6,814	3.11×10^{-5}	6.09×10^{-3}

LAGER.

First, we tune our LAE selection to match what LAGER has observed in the published 195 $z = 6.9$ LAEs utilized in the 4-field LAGER luminosity function (Wold *et al.*, 2021) with an average limiting luminosity of $\log_{10} L_{\text{Ly}\alpha} > 42.7$ erg s^{-1} and equivalent width (EW) threshold of $EW_{\text{Ly}\alpha} \gtrsim 10$Å. For such narrow imaging (approximately 30 cMpc for the LAGER narrowband), the flux-limited observations of LAEs mimics a line luminosity cut that would create a volume-limited sample. LAEs are known to have a possible duty cycle, meaning not all galaxies capable of emitting Lyα will be observed in the line. We therefore apply a duty cycle of 12.5% to the Jensen *et al.* (2014) simulations, following Kovač *et al.* (2007), and find excellent agreement between the observed LAE number density in the current 4 fields of LAGER (Wold *et al.*, 2021) and the $\langle x_i \rangle_m = 0.73$ simulation (closest to the derived limit of $\langle x_{i,\text{LAGER}} \rangle > 0.67$). We also confirm the simulations show great consistency with the observed number and surface densities of LAEs across the current 4 fields of LAGER, $\mathcal{N}_{\text{LAGER}} = 2.86 \times 10^{-5}$ Mpc^{-3} and $\Sigma_{\text{LAGER}} = 4.74 \times 10^{-3}$ arcmin^{-2}.
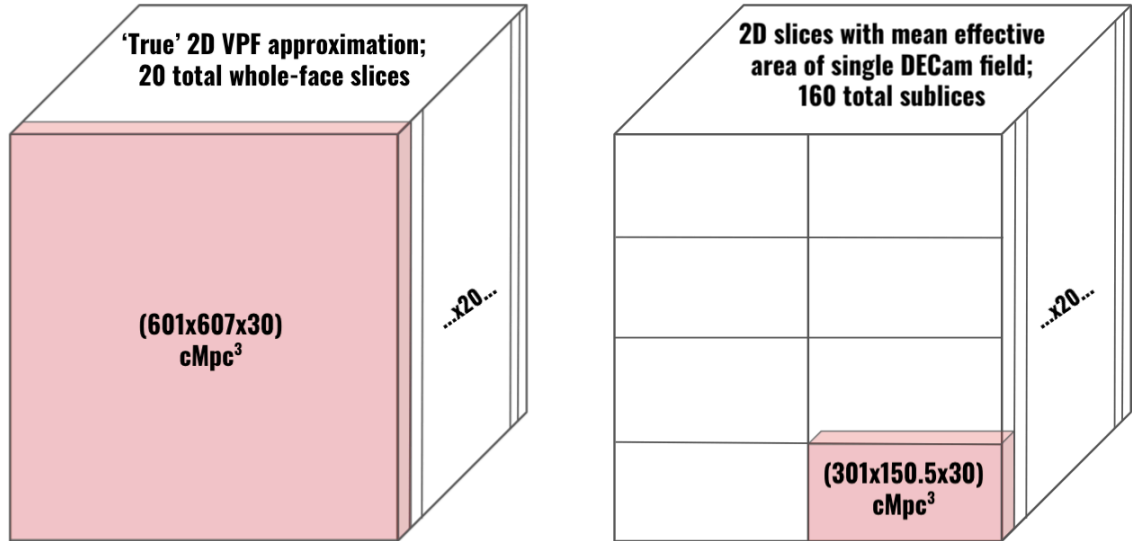
Figure 3.1: A Schematic Showing How the Jensen *et al.* (2014) Simulations Are Sliced To Explore How Reionization Constraints Change with Area Using the VPF in LAGER.

The number densities of the simulated LAEs range from 1.12×10^{-5} (30% ionized) to 3.11×10^{-5} (95% ionized) Mpc^{-3}. Table 3.1 details the number and densities of LAEs in the different simulations we use after the our selection.

Complete Table 3.1 caption: The mass-averaged ionized hydrogen fraction in the IGM of each simulation is $\langle x_i \rangle_m$; the corresponding volume-averaged ionized hydrogen fraction in the IGM (from Figure 2 of Jensen *et al.* 2013) is $\langle x_i \rangle_v$; N$_{\text{total}}$ is the total number of LAEs in each simulation before any selection; and N$_{\text{selected}}$ is the number of LAEs after the LAGER luminosity cut of $\log_{10}(\text{L}_{\text{cut}})$=42.7 erg s^{-1}. Wold *et al.* (2021) conservatively conclude, based on their analysis of the 4-field LAGER luminosity function with 195 LAEs, that at z=6.9 the neutral hydrogen fraction $\langle x_{\text{HI}} \rangle_m < 0.33$, or in the formalism of the Jensen *et al.* (2014) simulations, $\langle x_{i,\text{LAGER}} \rangle_m > 0.67$. The volume density is calculated with the entire (602x607x600)

Mpc^3 simulation, and the surface density is averaged across the 20 narrowband-thin 'slices' of (602x607x30) Mpc^3.

We now briefly summarize the observations LAGER has already completed and analyzed to motivate our slicing and sampling of the Jensen *et al.* (2014) simulations. Wold *et al.* (2021) detect their $z = 6.9$ LAEs in a total area of 10.19 deg^2 across the WIDE12 (3.24 deg^2), GAMA15A (2.91 deg^2), COSMOS (1.90 deg^2), and CDFS (2.14 deg^2) fields, or a total effective area of 7.6 deg^2. At $z = 6.9$, a cube face of (601x607 cMpc)^2 of the Jensen *et al.* (2014) simulations covers an area of 15.5 deg^2, allowing us to predict what the full 8 fields of LAGER may image if the same effective areas stand for future fields. We clarify here that although each DECam image observes 3.3 deg^2, each field in the Wold *et al.* (2021) LAE $z = 6.9$ luminosity function analysis averages 2.55 deg^2 due to a combination of bright foreground stars, limited ancillary data for LAE selection, and gaps from not dithering the detector in very early stages of the LAGER observations. We choose here to conservatively assume the future LAGER fields will cover at least similar effective areas, corresponding to the (301x150.5x30) cMpc^3 or $\sim$2 deg^2 slices. Finally, we emphasize that this work with the VPF and the Jensen *et al.* (2014) simulations is an independent assessment of the precision LAGER may find in its reionization constraints using clustering. This work was done independently to the work of Wold *et al.* (2021), had mostly taken shape before their work was completed, and is at its core agnostic to their results. We include their results here for context, observational grounding for part of the experiment set up, and to demonstrate that the predicted area coverage of LAGER was indeed met by the team's observations.

Figure 3.1 illustrates how we slice the Jensen *et al.* (2014) simulations for these LAGER tests. To get a sense of what the 'true' 2D VPF would be for the LAEs at this redshift and under the given ionization fraction, we create 20 slices of (602x607x30

cMpc)^3. To test what ionization fractions one, four, or eight sampled LAGER fields might probe with the VPF, we create 160 independent slices of (300x150x30) Mpc^3, corresponding to the average effective area of one LAGER imaging according to the 4 field analysis of Wold *et al.* (2021), and the approximate distance covered by the N964 filter. Once creating these small slices, we test what constraints the VPF of LAGER will be able to give for reionization for a single DECam imaging, the 4 fields that have been imaged and processed as of the writing of this work, and the total 8 fields planned for LAGER. Our choice of slices also incorporates the effects of cosmic variance on the VPF: in applying the cosmic variance calculator of Trenti and Stiavelli (2008) to the individual DECam-field-sized slices slices, we find a total fractional error of $\sim 24 - 30\%$, which is encompassed in the range of LAE counts we measure in the slices.

Complete Figure 3.1 caption: A schematic showing how we slice the Jensen *et al.* (2014) simulations in order to explore how the precision of reionization constraints changes with area using the VPF in LAGER. To approximate the 'true' 2D VPF value for the LAEs, we create 20 full-face slices of (601x607x30 cMpc)^3. This area corresponds to what the 8 currently planned fields of LAGER would likely cover for z=6.9 LAEs, were they all connected. This approach allows us to average over cosmic variance within the scale of the simulations, as well as include the signal of any large voids what might be split by the RA and Dec partitions we chose. To test what LAGER will measure in individual fields and surveys of different sizes, we create 160 subslices of (300x150.5x30 cMpc)^3. This corresponds to the approximate mean effective area across a single LAGER field (~ 2 deg^2), as observed and analyzed in Wold *et al.* (2021). In Sections 3.4, we randomly sample one, four, or eight subslices to predict the behavior and reionization constraints of the VPF once it is measured in the LAGER LAEs.

3.3 Constraining Reionization with the Void Probability Function

3.3.1 The Void Probability Function as a Probe

Neutral hydrogen acts like a fog for Lyman-α emission, meaning that LAEs during the epoch of reionization inscribe the amount and distribution of neutral hydrogen in the nearby IGM. The effect of reionization on LAEs can be subtle and measuring it quite model-dependent, so constraining this effect in various ways allows for stronger scientific consensus. Different clustering tools probe different aspects of galaxies' large scale structure and vary in performance at different scales. Several studies have been done with the angular correlation function (ACF) of narrowband-selected LAEs during reionization (Ouchi *et al.* 2010; Sobacchi and Mesinger 2015b; Ouchi *et al.* 2018; Gangolli *et al.* 2021) to constraint reionization with LAEs. However, small samples sizes, the resulting Poisson noise in the ACF, and the requirement of several distance scales to derive a correlation length can lead to significant levels of uncertainty in this vital constraint on reionization.

As the field gathers and awaits larger samples of observed LAEs during reionization to more finely measure the clustering signals caused by inhomogeneous reionization, these effects have been mostly explored in simulated LAEs. Jensen *et al.* (2013) explored the two-point correlation function in their early simulations, and later showed that count-in-cells showed a difference between ionization fractions under a single number density cut in Jensen *et al.* (2014). Recently, Gangolli *et al.* (2021) compared how several statistics, including the two-dimensional VPF, are able to constrain late reionization models and probe the effect of cosmic variance in mock LAE surveys made to mimic a SILVERRUSH-like survey at $z = 5.7$ and 6.6 (Ouchi *et al.* 2018; Konno *et al.* 2018). They found that the VPF is more sensitive than the angular correlation function when testing very late-stage ionization models under a

simple χ^2 analysis, even when incorporating high contamination fractions.

The Void Probability Function (VPF) can complement the ACF analysis by analyzing the volume-averaged clustering instead, and the guidelines of Perez *et al.* (2021) can help refine the scales of surveys and their analyses, as well as probe uncertainties in different ways. The VPF measures the probability of a given region being devoid of galaxies. It can also be thought of as a volume-averaged zero-point correlation function, and can connect to higher order correlation functions under hierarchical scaling frameworks. The VPF often complements count-in-cells (CiC) analyses, especially those that study the underlying physics behind hierarchical scaling (e.g. Conroy *et al.* 2005). Like CiC, the VPF is known theoretically for a given number density $\mathcal{N}$: $P_N = ((\mathcal{N}V)^N/N!)\exp(\mathcal{N}V)$, where $N = 0$ for the VPF so $P_0 = \mathrm{VPF} = \exp(\mathcal{N}V)$. The VPF can also give meaningful results at few distance scales, rather than the several the correlation function requires to confidently measure a power law.

The VPF is faster to calculate per-capita than CiC, as it only focuses on cells with zero galaxies, and as is motivated in Perez *et al.* (2021), has more intuitive errors and limits than CiC. Perez *et al.* (2021) also contend that for applications that do not require a correlation length (with the standard two-point correlation function) or an analysis of hierarchical scaling (with count-in-cells)–such as detecting an excess of LAE clustering due to inhomogeneous reionization–the speedy VPF and its intuitive bounds and errors might be a preferable option. In this work predicting what constraints for reionization LAGER may give with the VPF, we use both self-written algorithms developed in Perez *et al.* (2021) as well as the incredibly fast k-nearest neighbor method introduced in Banerjee and Abel (2020) to measure the VPF.

3.3.2 When Can We Trust a VPF Measurement?

A key goal of this work is to use the VPF to understand what precision in reionization constraints the narrowband survey of LAGER will be able to find as a function of area. Perez *et al.* (2021) derived conservative guidelines for when a VPF measurement is reliable, as well as minimum requirements of survey sizes to meet these guidelines. We apply these guidelines to the case of the LAGER survey below, and examine the VPF and its ability to to constrain the ionization fraction in detail in §3.4. Additionally, future reionization studies can use these guidelines and our analysis as a starting point to prepare for LAE surveys that will measure clustering.

A minimum radius for the VPF is simply the average distance between two random points, and relies on the density of the sample. Any smaller empty test spheres may not be true voids in the sample, but perhaps consequences of resolution. The maximum radius in this framework depends on the entire survey volume and the desired precision of the VPF. In order to measure $\log_{10}(P_0)$ to a given $-\alpha$ value and guarantee the level of precision of the VPF within $\pm 10^{-\alpha}$, the survey volume must be able to contain 10^α independent sub-volumes of the given radius. By considering the survey size and density, we can derive the number of galaxies needed for to measure the VPF to a given precision across a given dynamic range d. For the 2D VPF, these guidelines use survey area A and surface density Σ:

$$R_{\mathrm{min}} = \sqrt{\frac{1}{\Sigma\pi}}; \quad R_{\mathrm{max}} = \sqrt{\frac{A}{\pi 10^\alpha}}; \quad d = \frac{R_{\mathrm{max}}}{R_{\mathrm{min}}} = \sqrt{\frac{\Sigma A}{10^\alpha}}; \quad N_{\mathrm{total}} = \Sigma \times A = d^2 10^\alpha$$

$$(3.1)$$

The observed number density of LAEs is 2.86×10^{-5} Mpc^{-3} across the first 4 LAGER fields (Wold *et al.*, 2021), where each DECam field images about 3.3 deg^2 and the effective areas the LAEs cover is about 2.5 deg^2 per field. Once converted

to a surface density of 17.1 LAEs deg^{-2}, this yields a conservative minimum radius of approximately 8.2 arcminutes. For a single LAGER field of approximately 2.55 deg^2, the maximum radius to measure to $\log_{10}(P_0) = -1.5$ is 9.5 arcminutes. For the four-field effective area of 10.2 deg^2, the maximum radius to measure to the more precise $\log_{10}(P_0) = -2$ is 11 arcminutes. Finally, for the full-face simulation area of 15.5 deg^2, the maximum radius to measure to $\log_{10}(P_0) = -2$ is 13.3 arcminutes. Based on the VPF measurements in Figures 3.2-3.4 (described completely in and 3.4), values of $\alpha_{2\text{D}} > -1.5$ for the VPF easily encompass the large differences across ionized fractions and remain within our recommended guidelines.

However, we can additionally leverage our results in this work to refine the minimum VPF radius in a more practical way. For example, one could say that the minimum radius to measure the VPF is when the measurement is statistically distinct from the VPF of an unclustered distribution. For example, the 2D VPF of the full-face slices in Figure 3.2 begin to lie at least 3σ away from the Poisson curve near at least 2 arcminutes (if using the standard error pictured in Figure 2) or more realistically near 5 arcminutes (if instead using just the standard deviation). Therefore, we can update our conservative lower radius limit of the full 8-field LAGER (equal in area to one of the 20 slices) VPF measurement to be 5 arcminutes.

Complete Figure 3.2 caption: The 2D VPF of LAGER-like simulation slices of Jensen *et al.* (2014) at (30, 50, 73, 95) percent ionized hydrogen fractions in (purple, blue, yellow, red) respectively. We assume a Lyα luminosity cut of $\log_{10}\text{L}_{\text{Ly}\alpha} > 42.7$ erg s^{-1}, corresponding to the approximate limiting line flux of the observed LAGER LAEs. The (602x607x600) Mpc^3 simulations have been divided into 20 slices of (602x607x30) Mpc^3, corresponding to the approximate depth of the LAGER narrowband at $z_{\text{Ly}\alpha} \sim 6.9$ and effective area projected for the 8 planned LAGER fields. With these large slices, we can approximate the 'true' 2D VPF for the simulated

LAEs at each ionization fraction (by covering large enough areas to minimize cosmic variance), and later compare to the 2D VPF we measure with LAGER DECam-sized subslices of these simulations. We plot the mean 2D VPF across the 20 slices as colored circles and shade the 3σ standard error (3 standard deviation divided by the square root of the number of samples). Different ionization fractions show varying number densities and very different VPFs, both from the effect of decreasing neutral fractions and also evolution in the host dark matter halos and galaxies themselves over cosmic history.

3.4 Constraining Global Neutral Fractions with the VPF from LAGER

In this section, we ask and answer: what area is needed to make precise constraints on reionization for the LAGER survey with LAE clustering? Specifically, how distinguishable are the different ionization fractions using the VPF for narrowband-detected volume limited samples as a function of survey volume? To do this, we use the conservative guidelines of Perez *et al.* (2021) to identify which radii will give trustworthy VPF measurements, and focus upon the behavior of the VPF at the smallest and largest trusted scales for when we pick and sample one, four, or eight LAGER-like fields.

As discussed in §3.3, the clustering of LAEs is affected in its extent and distribution by reionization. Instigating a single luminosity cut across all ionization fraction volumes mimics how narrowband-detected LAEs at different redshifts and neutral fractions show a clear drop in their number density and increase in their clustering. This closely complements analyses of the Lyα luminosity function as it evolves through the epoch of reionization, and will offer a constraint of the neutral fraction of the IGM under the same effect with different models and simulations, and utilizing the same data with very little modification.

72
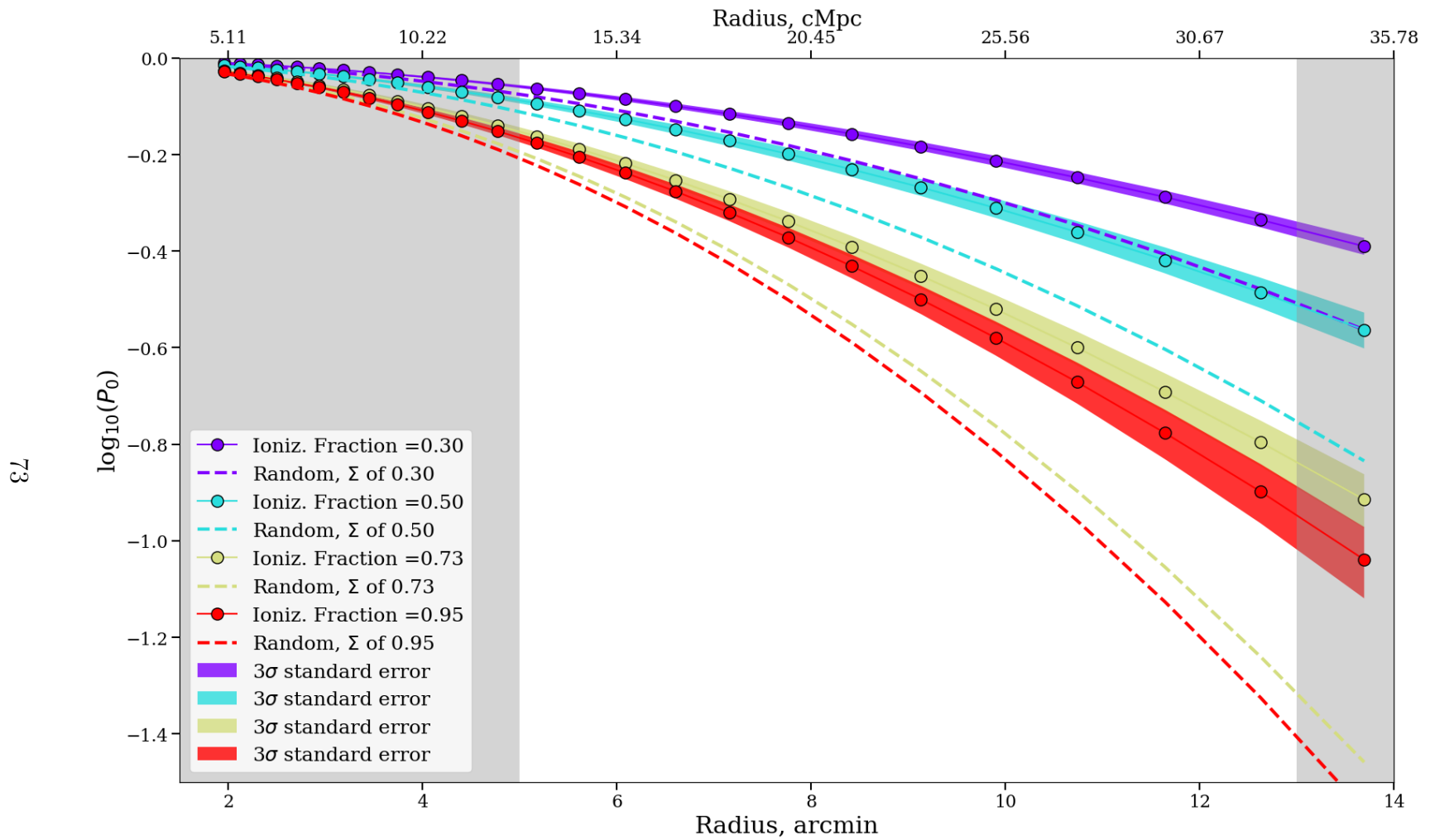
Figure 3.2: The 2D VPF of LAGER-like Simulation Slices of Jensen *et al.* (2014) at (30, 50, 73, 95) Percent Ionized Hydrogen Fractions.
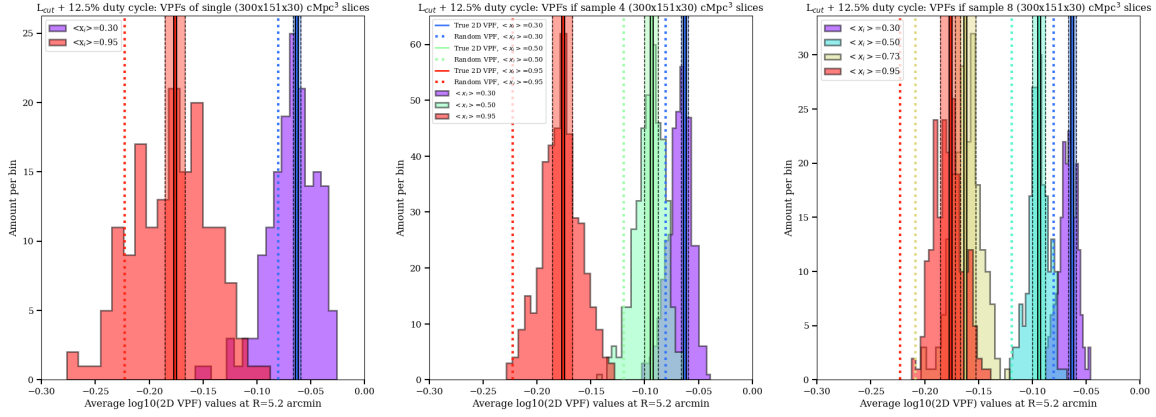
Figure 3.3: The distribution of the 2D VPF measured at 5.2 arcmin (13.23 Mpc) for: all 160 individual DECam-sized subslices (left), the average when randomly choosing four subslices (center) four hundred times, and the average when choosing eight subslices (right) two hundred times. The solid lines are the 2D VPF of the full-face volumes for the given ionization fraction in Figure 3.2, and serve as our approximation of the 'true' VPF value at this radius. The 3σ standard error from of the full-face 2D VPF are the colored shaded regions enclosed with thin black dotted lines. The VPFs at this radius for completely unclustered Poisson distribution at the relevant surface densities (averaged across the 20 whole-face slices) are approximately {-0.08, -0.13, -0.23} for {0.30, 0.50, 0.73/0.95}.
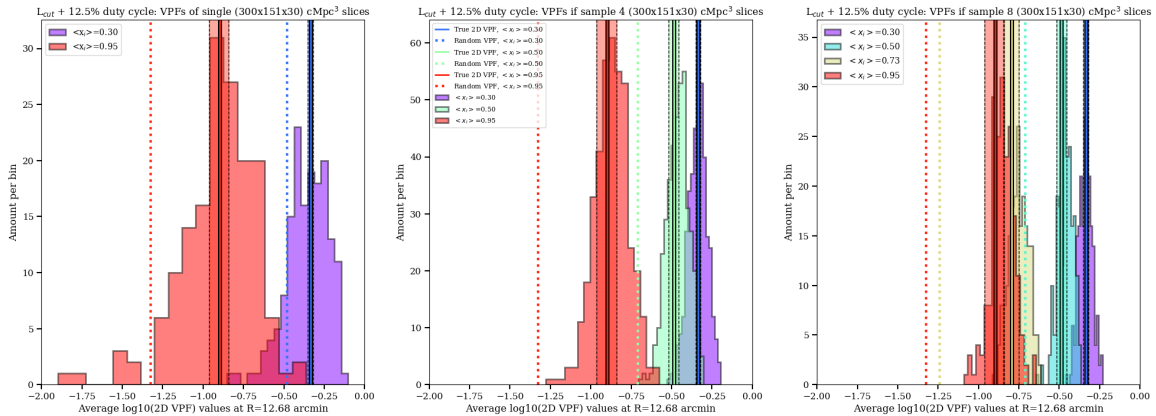
Figure 3.4: Like Figure 3.3, but at the maximum radius of 12.68 arcmin (32.27 cMpc).

Table 3.2: Details of the VPF Distributions in Figures 3.3 and 3.4. Once cutting up the Jensen *et al.* (2014) simulations into 160 LAGER DECam-sized slices of (301x150.5x30) Mpc^3, we measure VPF at the minimum and maximum scale for LAGER across all individual slices, and then when randomly picking and averaging four or eight slices. We find the distributions are well described by Gaussian functions, and we list the mean $\log_{10}$VPF value and full width at half max for each distribution. We also calculate the approximate overlap between the simulations' curves for each sampling.

Histogram selection					$R_{min} = 5.2$ arcmin			$R_{max} = 12.7$ arcmin	
Sample size	$\langle x_i \rangle_m$	Peak VPF	FWHM	% overlap		Peak VPF	FWHM	% overlap	
All 160 single slices	0.30	-0.064	0.22	18% 0.30-0.95		-0.34	0.13	40% 0.30-0.95	
...	0.95	-0.18	0.039	...		-0.87	0.19	...	
Pick 4, 400 times	0.30	-0.066	0.013	0% 0.30-0.95		-0.33	0.057	0% 0.30-0.95	
...	0.50	-0.094	0.013	72% 0.30-0.50		-0.46	0.07	74% 0.30-0.50	
...	0.95	-0.177	0.018	3% 0.50-0.95		-0.85	0.102	6% 0.50-0.95	
Pick 8, 200 times	0.30	-0.068	0.0075	0% 0.30-0.73		-0.34	0.04	0% 0.30-0.73	
...	0.50	-0.095	0.01	19% 0.30-0.50		-0.45	0.05	54% 0.30-0.50	
...	0.73	-0.16	0.012	0% 0.50-0.73		-0.74	0.06	0% 0.50-0.73	
...	0.95	-0.18	0.013	80% 0.73-0.95		-0.86	0.07	80% 0.73-0.95	

As derived in §3.3.2, a *single* DECam field of 3.3 deg^2 at the observed LAE surface density of 4.74×10^{-3} arcmin^{-2} will yield the conservative minimum and maximum radii limits of about 8 and 10 arcminutes (respectively). In recognizing that real galaxies will be clustered enough to be statistically distinct from random at smaller radii, we choose to focus on the 2D VPF measured at $R = 5.2$ arcminutes. At this radius, the mean VPF value of the 20 full-face slices is at least 3 standard deviations away from the predicted value for a random distribution for all the simulations. In comoving units, this corresponds to 13.3 cMpc, within our 25 radii spaced logarithmically between 5 and 35 Mpc. To contrast this, we examine the VPF at 12.68 arcminutes (32.27 cMpc), near the largest radius that our conservative guidelines suggest for the fully planned 8 LAGER fields (assuming a similar yield of effective area as seen in Wold *et al.* 2021, approximately 15.5 deg^2).

In Figure 3.2, we show the 2D VPF under a single luminosity cut for the 20 whole-face slices of (602x607x30) cMpc^3. This mimics the 'true' value and inherent variance of the 2D VPF for narrowband-selected LAEs at $z = 6.9$, given the simulation assumptions and models. The colored circles are the mean VPF across the 20 slices, the shading is the 3σ standard error (where 1σ standard error is one standard deviation divided by the square root of the number of slices), and the dashed lines are the theoretical 2D VPF for an unclustered distribution at the surface density of the given simulation. Each ionization fraction simulation is strongly clustered compared to its random curve; and promisingly for our analysis, each ionization fraction's 2D VPF curves is statistically distinct from the others. As we will find when examining the spread of the 2D VPF across individual DECam-sized slices, the large separations allows for us to constrain mostly ionized vs. mostly neutral given the Jensen *et al.* (2014) model assumptions with the four currently observed LAGER fields.

To probe how survey area affects the constraints that LAGER might give for

reionization using the 2D VPF of LAEs, we leverage the size of the Jensen *et al.* (2014) simulations to examine the behavior of the 2D VPF across 160 subslices similar to individual DECam imagings. In Figures 3.3 and 3.4, we plot the measured 2D VPFs under one Lyα luminosity cut (essentially one flux limit in our narrowband).The leftmost subplot shows the distribution of 2D VPFs for all 160 (300x150x30) Mpc^3 individual subslices (left); the center shows the distribution when we randomly "pick 4" subslices and average them four hundred times; and the left shows the distribution when we randomly "pick 8" subslices and average them two hundred times. The "pick 4" and "pick 8" strategy mimics how the LAGER team will approach the clustering measurements of the four current and eight total planned fields. In Table 3.2, we list for Figures 3.3 and 3.4 the approximate central $\log_{10}$VPF value of each distribution, the full width at half maximum, and approximate percent overlap of each simulation's distribution with its neighbors'.

Complete Figure 3.4 caption: What constraints on ionization fraction can the VPF give for various survey areas of LAGER, at a small VPF scale? We plot the distribution of the 2D VPF measured at 5.2 arcmin (13.23 Mpc) for: all 160 individual DECam-sized subslices (left), the average when randomly choosing four subslices (center) four hundred times, and the average when choosing eight subslices (right) two hundred times. Sampling more subslices will lead to averages that are more narrowly distributed around the true average value. The solid lines are the 2D VPF of the full-face volumes for the given ionization fraction in Figure 3.2, and serve as our approximation of the 'true' VPF value at this radius. The 3σ standard error from of the full-face 2D VPF are the colored shaded regions enclosed with thin black dotted lines. The VPFs at this radius for completely unclustered Poisson distribution at the relevant surface densities (averaged across the 20 whole-face slices) are approximately {-0.08, -0.13, -0.23} for {0.30, 0.50, 0.73/0.95}. The left panel shows that a single

LAGER field could constrain between 30% or 95% ionized universe; the center shows that four averaged fields might additionally distinguish 50% ionized around -0.15< $\log_{10}P_0$ < -0.1; and the right shows that eight averaged fields more clearly distinguish 50% ionized, and may in rare cases also separate 73% from 95% ionized fraction where -0.16 < $\log_{10}P_0$ < -0.12.

Complete Figure 3.3 caption: Like Figure 3.3, but measured at the maximum radius of 12.68 arcmin (32.27 cMpc). The VPFs at this radius for completely unclustered Poisson distribution at the relevant surface densities (averaged across the 20 whole-face slices) are approximately {-0.5, -0.75, -1.25, -1.3} for {0.30, 0.50, 0.73, 0.95}, and shown in colored dotted lines. The 3σ standard error from of the full-face 2D VPF are the colored shaded regions enclosed with thin black dotted lines. Like Figure 3.3, these distributions show that the VPF measured at large scales can also constrain whether an individual LAGER field is more likely to be 30% or 95% ionized, and in some cases statistically distinguish a 50% ionization fraction under this model as well. Individual VPF measurements become less correlated the more separated they are in distance, allowing a verification of constraints at small distance scales.

Complete Table 3.2 caption: Details of the VPF distributions in Figures 3.3 and 3.4. Once cutting up the Jensen *et al.* (2014) simulations into 160 LAGER DECam-sized slices of (301x150.5x30) Mpc^3, we measure VPF at the minimum and maximum scale for LAGER across all individual slices, and then when randomly picking and averaging four or eight slices. We find the distributions are well described by Gaussian functions, and we list the mean $\log_{10}$VPF value and full width at half max for each distribution. We also calculate the approximate overlap between the simulations' curves for each sampling, finding what percentage of a given simulation's VPF are measured at the same value in any another simulation, and accounting for variations due to binning and the sampling.

When assessing the distribution of the 2D VPF for individual subslices, the 30% and 95% simulations are quite distinct. As seen in Figure 3.3 and quantified in Table 3.2, the simulations overlap at their only tails around $\log_{10}P_0 \approx -0.13$. LAEs in a given LAGER field are more likely to be mostly ionized if their $\log_{10}P_0$ is measured to be less than -0.13, and mostly neutral if their $\log_{10}P_0$ is greater than -0.13. When we randomly sample and average four fields (aka. "pick 4"), the center panels of Figures 3.3 and 3.4 show how the 30 and 95 % ionized fraction simulations completely separate with no overlap, allowing one to confidently determine that the $z = 6.9$ universe is consistent with being either almost entirely ionized or mostly neutral with the VPF of four LAGER fields. Additionally, the 50% simulation fills much of the gap between them, even allowing one to tell apart 30 vs. 50% ionized fractions near $\log_{10}P_0 = -0.1$ (-0.5 in Figure 3.4). With eight fields sampled (aka. "pick 8"), as pictured on the right side of Figure 3.3, the distributions will become narrower and allow more specific ionization fraction measurements by the Jensen *et al.* (2013) models, differentiating even 73% and 95% ionized fractions for rare samplings. For example, near $\log_{10}$VPF=-0.14 in Figure 3.3 or $\log_{10}$VPF=-0.6 in Figure 3.4, the 95% distribution nearly never samples but the 73% distribution often does.

In Figure 3.4, we examine the VPF distributions at 12.68 arcmin (32.27 cMpc), the largest distance scale our covered area allows us to measure. This corresponds to the 13 arcminutes allowed by the projected 15.5 deg^2 for the full 8 fields planned for LAGER for a VPF sensitivity of $\log_{10}(P_0) = -\alpha = -2$. The correlations between individual VPF measurements will decrease the more separated they are in distance scales, and VPF measurements are least correlated at small scales and most correlated at the largest scales (Gangolli *et al.* 2021 and Gangolli, private communication). This allows us to confirm the constraint on the ionized hydrogen fraction with the VPF from the smallest distance scale, while minimizing the correlation between the

clustering measurements.

As with the measurement at our smallest distance scale, at this largest distance scale, the separation between 30% and 95% ionized is significant with even a single DECam field. The 30% and 95% ionized simulations can be further distinguished from 50% with 4 fields, and in rare cases also 73% ionized with all 8 planned fields of LAGER. We stress again that we have somewhat pessimistically assumed the next 4 fields imaged by LAGER will cover the same effective area of the completed 4 fields. If the 4 fields currently undergoing imaging and analysis all have effective areas of more than 3 deg^2 (as did WIDE12), the increase in both the number of LAEs detected and covered area will mean the VPF can be measured more precisely to larger dynamic range, with clearer understanding of cosmic variance, and further constrain the IGM ionization fraction at $z = 6.9$.

Finally, let us further focus on the possibility of constraining between the most ionized fractions, relevant to various studies attempting to constrain late stage reionization. As mentioned before, we find that the 8 planned fields of LAGER will rarely distinguish 73% from 95% ionized. This comes from examining the third panels of Figures 3.3 and 3.4: the 73% from 95% distributions overlap significantly (as Table 3.2 notes pessimistically, about 80% of these distributions overlap in any amount). Therefore, the 8-field VPF measurement would need to lie at the far edges of either distribution (i.e. be a particularly clustered 73% value, or particularly *un*clustered 95% value). We do see that the full uninterrupted 'full-face' $\sim$24 deg^2 measurement of the VPF separates the 73% and 95% distributions to 3σ, leading to the question: how many DECam fields' worth of imagining are needed to reach a similarly narrow distribution? As Figures 3.3 and 3.4 show, 8 fields is not too far from that yet; therefore, perhaps 12 fields of LAGER, for approximately 30 deg^2, could further distinguish 73 and 95 % ionized. This agrees with Gangolli *et al.* (2021)'s analysis with

the $z = 5.7$ LAE VPF in e.g. Figure 6, where the VPF "1σ dipersion from cosmic variance in the VPF (assumed) a futuristic... SILVERRUSH-like survey with roughly double the size" (approximately 40 deg^2). If this is true, the VPF of a larger LAGER survey would further refine Wold *et al.* (2021)'s constraint of $\langle x_{i,\text{LAGER}} \rangle > 0.67$ or $\langle x_{\text{HI, LAGER}} \rangle < 0.33$.

3.5   Summary and Conclusions

We leverage the uncommon VPF statistic to quantify how precise constraints on reionization from clustering in the Lyman-α Galaxies in the Epoch of Reionization (LAGER) narrowband survey may be. Specifically, we probe how distinguishable different ionization fractions are when using the VPF for narrowband-detected volume limited samples as a function of survey volume. We use the simulations of LAEs throughout reionization from Jensen *et al.* (2014), and leverage the large (600 Mpc)^3 volumes to statistically probe what the VPF of LAGER might measure for one, four, or eight DECam imagings of LAEs at $z = 6.9$. We apply a Lyα line luminosity cut of $\log_{10}\text{L}_{\text{Ly}\alpha} > 42.7$ erg s^{-1}, as measured in the observed LAEs of Wold *et al.* (2021). Finally, we create volumes of LAEs within an IGM whose hydrogen is 30, 50, 73, or 95% ionized, slicing each to the width that the LAGER N964 filter spans at $z = 6.9$, approximately 30 cMpc.

We measure the VPF on the 20 whole-face (602x607x30) Mpc^3 slices as the 'true' VPF with minimal cosmic variance, given the Jensen *et al.* (2013) models and the LAGER flux detection limits, in Figure 3.2. We then create slices of (301x150.5x30) Mpc^3 to serve as samples of what a single DECam imaging might measure, based on the effective areas across the four fully-imaged LAGER fields analyzed in Wold *et al.* (2021). To understand what constraints for reionization the cumulative LAGER survey might give with the VPF, we randomly sample four (eight) subslices four (two)

hundred times, creating distributions of what the VPF measurement might look like for the four already analyzed LAGER fields and with the additional four fields in progress (Table 3.2). We examine the VPF distributions near the smallest distance scale given the observed LAE surface density (near 5 arcmin), and the maximum radius implied by the full 8 field LAGER effective area (near 13.5 arcmin).

We note that all the Jensen *et al.* (2014) simulations' morphology for reionization comes from the single ionization scenario that was simulated, and that the shapes of neutral and ionized regions will change with other assumptions for the sources and sinks of Lyα emission. Our results are model-dependent upon the single scenario run, and offer an additional comparison to models used for other Lyα studies. Additionally, the large size of the simulations allows us to answer how much volume is needed to constrain ionization fraction using the VPF, especially for the unique geometry of LAGER. Future work, for example, will extend this type of analysis to determining what constraints on models for the pacing of ionization the Roman Space Telescope will be able to give for deep and wide surveys of LAEs between $7.2 < z < 14$ (Perez et al., in prep).

Under the Jensen *et al.* (2013) models and with the Jensen *et al.* (2014) simulations, we find that even a single DECam field might be able to discriminate between mostly neutral or mostly ionized based on the VPF distributions in Figures 3.3 and 3.4. Sampling and averaging four fields allows the distinction between 30, 50, and 95 percent ionized, and utilizing eight might even further distinguish between 73 and 95 percent ionized. An expanded LAGER survey of ten or more fields, with 30+ deg^2 would confidently distinguish 73 and 95 percent ionized, and perhaps better refine Wold *et al.* (2021)'s constraint of $\langle x_{i,\text{LAGER}} \rangle > 0.67$ or $\langle x_{\text{HI, LAGER}} \rangle < 0.33$. Finally, as proposed in McQuinn *et al.* (2007), the VPF of a fully ionized sample under the same luminosity selection (e.g. narrowband LAEs at z=5.7 or below) can be compared to

the VPF of the LAGER $z = 6.9$ LAEs to further constrain reionization with Jensen *et al.* (2013) model. Future work in the LAGER collaboration will examine the VPF and other clustering statistics across the LAEs of all completed and ongoing fields, and be informed by the precision and constraints predicted by our work.

Chapter 4

EVOLUTION OF THE VOID PROBABILITY FUNCTION OF LYMAN-ALPHA EMITTERS THROUGHOUT REIONIZATION, AS OBSERVED WITH THE ROMAN SPACE TELESCOPE

Lucia A. Perez[1], Sangeeta Malhotra[2], James E. Rhoads[2], Isak Wold[2]

[1] Arizona State University School of Earth and Space Exploration, Tempe, AZ

[2]NASA Goddard Space Flight Center, Greenbelt, MD

A paper based on this chapter has been submitted to the Astrophysical Journal; please see the refereed publications for updated conclusions based on peer review.

ABSTRACT

We use large simulations of Lyman-Alpha Emitters with different fractions of ionized intergalactic medium to quantify the clustering of Lyα emitters as measured by the Void Probability function (VPF), and how it evolves under different ionization scenarios. We quantify how well we might be able to distinguish between these scenarios with a deep spectroscopic survey using the future Nancy Grace Roman Space Telescope. Since Roman will be able to do blind spectroscopic surveys of Lyα emitters continuously between $7 < z < 12$ to sensitivities of at least 10^{-17} erg sec^{-1} over wide field of view, we will measure the epoch of reionization as well as the pace of ionization of the intergalactic medium (IGM). We compare deep Roman surveys covering roughly 1, 4 and 16 deg^2, and constraints on reionization the VPF may find. A survey of 1 deg^2 would distinguish between very late reionization and early reionization to 3σ near $z = 7.7$ with the VPF. The VPF of a 4 deg^2 survey can distinguish between slow vs. fast, and early vs. late, reionization at $> 3 - 4\sigma$ at several redshifts between $7 < z < 9$. However, a survey of 13-16 deg^2 would allow the VPF to give several robust constraints ($> 5 - 8\sigma$) across the epoch of reionization, and would yield a detailed history of the reionization of the IGM and its effect on Lyman-α Emitter clustering.

4.1 Introduction

The epoch of reionization (EoR) is the era when the earliest galaxies ionized the ultraviolet-opaque 'fog' of neutral hydrogen that filled the early universe. Reionization history is still not well constrained, as various probes have led to different conclusions about its timing, pace, and sources (McQuinn, 2016). Lyman-α emitters (LAEs), which are star-forming galaxies that strongly emit in the Lyα line, offer practical probes of the reionization process. Detecting LAEs by their line emission enables surveys of otherwise faint galaxies across a wide redshift range. Because the Lyα line is resonantly scattered, it is easily attenuated by any amount of neutral hydrogen, and its visibility is highly sensitive to the ionization state of intergalactic gas throughout the reionization process (as predicted by Miralda-Escudé 1998; Haiman and Spaans 1999).

A common method of constraining reionization with LAEs uses the Lyα luminosity function (LF; first implemented by Malhotra and Rhoads 2004b; e.g. Ouchi *et al.* 2010; Santos *et al.* 2016; Ouchi *et al.* 2018; Morales *et al.* 2021). This works by comparing the observed Lyα LF to that expected in a fully ionized medium (commonly based on the observed LF at a redshift where reionization is believed to be complete, e.g. $z = 5.7$, Ajiki *et al.* 2003; Ouchi *et al.* 2003; Hu *et al.* 2004).

LAE LF studies have reached various constraints for reionization. Some report suppression of the LAE LF over (part of) the range $5.7 < z < 7$, with inferred neutral fractions of order 20–40 percent (McQuinn *et al.* 2007; Kashikawa *et al.* 2006; Iye *et al.* 2006; Ouchi *et al.* 2010; Kashikawa *et al.* 2011; Konno *et al.* 2014, 2018), while others report upper bounds on neutral IGM (Malhotra and Rhoads 2004b, 2006), including some bounds that are tight enough (< 20–30 percent at $z = 7$) to challenge some of the suggested detections of neutral gas (Wold *et al.* 2022). LF suppression can have

causes beyond a partially reionized IGM, such as cosmic variance, the evolution of the halo mass function (Dijkstra *et al.*, 2007), and evolution in various factors that affect which galaxies strongly emit Lyα (e.g. Ota *et al.* 2008; Stark *et al.* 2010; Pentericci *et al.* 2011; Ono *et al.* 2012; Schenker *et al.* 2012; Endsley *et al.* 2021).

The clustering of EoR LAEs is independent from the evolution of their intrinsic LF, and provides an additional method of constraining reionization to help resolve these tensions (McQuinn *et al.*, 2007). Most constraints have used the angular two-point correlation function (Ouchi *et al.* 2010; Sobacchi and Mesinger 2015b; Ouchi *et al.* 2018; Gangolli *et al.* 2021) or related statistics (e.g. count-in-cells in Jensen *et al.* 2014). In this work, we focus on the *Void Probability Function*, a less common choice for EoR LAEs (Kashikawa *et al.* 2006; McQuinn *et al.* 2007) but one that may offer more sensitivity than the angular correlation function under some models (Gangolli *et al.*, 2021). The VPF measures clustering by measuring how likely a circle or sphere of a given size is to be empty in a galaxy sample. As the 0th moment of count-in-cells, or zero-point volume-averaged correlation function, it carries the signature of higher order correlation functions. Its simplicity can be used to derive guidelines for the required density and volume of surveys (Perez *et al.*, 2021).

In Perez et al. (2022a, submitted; henceforth Paper 1), we quantified what constraints the VPF could yield for reionization when applied to large-area narrowband searches for LAEs, such as the Lyman-Alpha Galaxies in the Epoch of Reionization (LAGER) narrowband survey. LAGER detects $z = 6.9$ LAEs using a narrowband filter centered at 9642Å with FWHM=92Å (corresponding to approximately 30 cMpc at $z = 6.9$), mounted on the 3.3 deg^2 field-of-view DECam on Cerro Tololo's 4-meter Blanco Telescope. So far, LAGER has yielded initial constraints of $\langle x_i \rangle_v > 0.67$ with its four-field LF (Wold *et al.* 2022, building off Hu *et al.* 2019), and has the observation and analysis of another four fields in progress. Using the Jensen *et al.* (2014)

simulations, Paper 1 predicts possible constraints for various iterations of LAGER with the VPF. We identify how well 1, 4, or 8 LAGER DECam fields distinguish different ionization fractions with the VPF, and lay out a framework and case for using the two-dimensional VPF together with the Jensen *et al.* (2014) simulations to constrain the ionized hydrogen fraction implied by LAE clustering.

While narrowband surveys like LAGER (Zheng *et al.* 2017; Hu *et al.* 2019, 2021; Wold *et al.* 2022) and SILVERRUSH (Ouchi *et al.* 2018; Konno *et al.* 2018) have been constraining the end of reionization with LAEs, the higher redshift EoR has been much more difficult to study on a large scale with LAEs. Ground-based infrared surveys are increasingly impractical at redshifts beyond $z = 6.9$ (where reionization is thought to be mostly complete), both due to atmospheric OH emission creating prohibitively bright sky foreground at most wavelengths, and a steep drop in silicon detector response at 1μm that restricts searches to instruments with comparatively small fields of view. Despite these challenges, focused ground-based searches for Lyα at $z = 7.7$ and beyond (Tilvi *et al.* 2020; Oesch *et al.* 2015; Zitrin *et al.* 2015) have yielded detections of Lyα emitters.

The Roman Space Telescope, NASA's next flagship mission set to launch in the mid-to-late 2020's, is an infrared telescope with a 2.4 meter Hubble-sized mirror, and a wide-field instrument with a 0.281 deg^2 field of view (200 times that of Hubble's WFC3-IR). Roman's wide-field instrument will have a slitless grism that is capable of capturing Lyα at $7.2 < z < 14$, as well as a lower-dispersion prism that will reach Lyα at lower redshifts ($z > 6$). Roman can carry out surveys of reionization-era LAE clustering that will notably refine our understanding of the EoR, giving definitive constraints for how and when reionization occurred.

In this work, we adapt our framework from Paper 1 to make predictions for LAE clustering observations with Roman. We quantify how Roman will constrain the tim-

ing and pace of reionization given its wide field, sensitivity, and continuous redshift coverage by combining three ingredients: a model for reionization in a $\sim (600\mathrm{cMpc})^3$ box that includes galaxy formation and radiative transfer of Lyα (Jensen $et~al.$, 2014); a set of reionization history models that we use to map between redshift and simulated neutral fraction; and instrumental sensitivity predictions based on detailed simulations of Roman grism data sets (Wold et al, in prep). We then calculate the expected VPF and its expected uncertainties for surveys covering ~ 1, ~ 4, and ~ 16 square degrees, at each of five neutral fractions ($\langle x_i \rangle_v = \{0.22,~0.40,~0.485,~0.66,~0.93\}$), in each of three reionization scenarios (gradual; rapid and early; rapid and late). We thereby address how large a clustering survey must be to distinguish between models for reionization history.

This paper is laid out as follows. In §4.2, we describe the theory and simulations that support this work: the Jensen $et~al.$ (2014) simulations of LAEs through reionization in §4.2.1; the models for reionization we compare in §4.2.2; our application of the Wold et al. (in prep) Roman-Lyα grism simulations in §4.2.3; how we create mock Roman LAE surveys from the above in §4.2.4; and finally, how we apply the VPF for reionization constraints in §4.2.5. In §4.3, we explore how the VPF evolves throughout the reionization history of the universe with different survey constructions, and how precisely Roman will be able to make out the epoch and pace of reionization. We focus on an ambitious 13-16 deg^2 survey in §4.3.1; smaller 1 and 4 deg^2 surveys in §4.3.2; and a focused $z \sim 7.75$ model comparison for the entire VPF(R) curve in §4.3.3. Appendix B shows in detail our VPF measurements. We conclude and summarize our results in §4.4.

4.2 Simulations and Methods

4.2.1 Large Simulations of LAEs in Partially Reionized IGM

For samples of LAEs at various ionization fractions, we use the Jensen *et al.* (2014, J14) simulations of LAEs, which include radiative transfer through various amounts of ionized intergalactic medium (IGM). The simulations exist for discrete *volume* ionized fractions of {22, 40, 48.5, 66, 78, 89, and 93} percent, corresponding to mass ionized hydrogen fractions of {30, 50, 58, 73, 83, 92, and 95} percent. We direct readers to Paper 1 for a more complete description of the J14 simulations and their properties for contexts similar to this work, but review relevant details for their assumptions for reionization (explained fully in Jensen *et al.* 2013).

The underlying reionization model of the J14 simulations can be summarized as an inside-out reionization history that began somewhat early and progressed at a moderate pace. They incorporated self-regulation as galaxies take part in reionization, turning off small sources of ionizing radiation once the IGM passed 10% ionization. The LAE catalogs at the highest ionization fractions ($\langle x_i \rangle_v > 0.6$) are consistent with the luminosity function of $z = 5.7$ LAEs (Ouchi *et al.*, 2010). The simulations model evolution of the halo mass function and galaxy luminosity function along with the IGM neutral fraction, so that differences in the galaxy population between different neutral fractions are automatically included in our analysis. Moreover, the intrinsic evolution of Lyα galaxies should be modest during the rapid phases of reionization. For example, $\Delta z = 0.5$ only corresponds to 65 Myr at $z = 7$.

Later in this work, we will redshift the LAE catalogs to reflect the redshifts different models for the EoR predict for each ionized IGM fraction (i.e. slow/fast, early/late reionization), and then consider what Roman will observe in the context of the VPF of LAEs. We will primarily focus on the $\langle x_i \rangle_v$={22, 40, 48.5, 66, 93} percent ion-

ized simulations. Most of these fractions are predicted to exist at redshifts where the Roman grism will observe Lyα under the models of reionization that we analyze. The most ionized volume is in particular a useful baseline of clustering, and can be compared to ground-based surveys as well as future studies with the Roman prism that will detect Lyα to $z > 6$.

4.2.2 Models for Reionization

We focus on a representative sample of models for the reionization history of the universe to guide our projections for Roman. Figure 4.1 shows the redshift vs. volume-averaged fraction of ionized hydrogen in the IGM, according to the reionization models of Finkelstein *et al.* (2019, F19), Naidu *et al.* (2020, N20), and Yung *et al.* (2020, Y20) (respectively, blue, red, and green). Figure 4.1 shows the discrete ionization fractions of the J14 simulations as dashed horizontal lines ranging from most neutral (purple) to most ionized (red).

These models sample a range of reionization histories, driven by different assumptions about the production of ionizing photons. These models are tuned to reproduce existing constraints for reionization, such as: measurements of the IGM temperature that indicate a mostly ionized IGM at $z \sim 6$ (e.g. Fan *et al.* 2006); analyses of measured Lyα line profiles that rule out a fully neutral universe at $z = 6.6$ (Haiman and Cen 2005; Ouchi *et al.* 2010; Rhoads *et al.* 2013); quasar Gunn and Peterson (1965) trough studies that find a fully ionized universe by $z = 5 - 6$ (e.g. Fan *et al.* 2006; McGreer *et al.* 2015); and measurements of the average Thomson scattering optical depth with cosmic microwave background (CMB) anisotropies (e.g. Dunkley *et al.* 2009; Planck Collaboration *et al.* 2016) that support higher-redshift extended reionization ($z > 7 - 8$). These models also focus on galaxies as the likeliest and dominant sources of reionization (e.g. Robertson *et al.* 2013), supported by the rarity of quasars
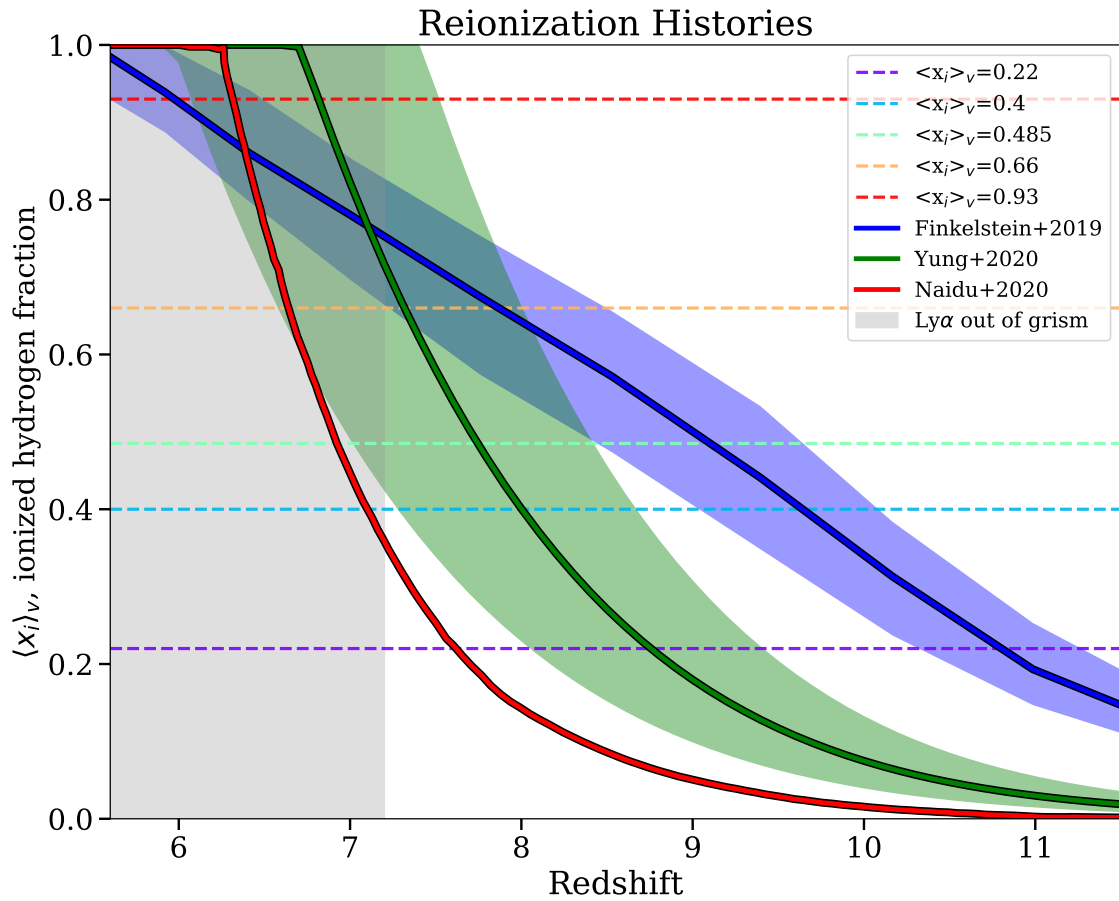
Figure 4.1: The predicted volume-averaged ionized hydrogen fraction by redshift according to several models of reionization: Finkelstein *et al.* (2019, F19) (blue), Yung *et al.* (2020, Y20) (green), and Naidu *et al.* (2020, N20) (red). We identify at which redshift the models (solid lines) predict the ionization fractions simulated by Jensen *et al.* (2014, J14) (dotted horizontal lines). The grey shaded region indicates the redshift (and relevant model information) the Roman grism cannot observe for Lyα.

beyond $z > 6$ and constraints on helium reionization (e.g. Madau and Haardt 2015). Next, we briefly describe the models for reionization that this work will compare.

F19 present a semi-empirical model of reionization whose core is a physically-motivated halo mass-dependent parametrization of escape fractions. In their model, the faintest galaxies in the UV collectively dominate the ionizing emissivity, leading to a reionization history that starts very early (80% volume-ionized at $z \sim 7$) and progresses at a smooth pace. They also model an AGN contribution to the end of reionization, making up one third of the budget at $z = 6$, and predict a flat star formation rate density at $z > 8$.

Y20 use end-to-end semi-analytic models with the goal of modeling in fine detail all ionizing sources. They use physically motivated relationships between dark matter halo formation histories and galaxy properties (including synthetic spectral energy distributions) to connect galaxy formation physics to the large-scale reionization history. They combine the Santa Cruz semi-analytic model for galaxy formation (Somerville and Primack 1999; Somerville *et al.* 2008, 2015) with an analytic reionization model (Madau *et al.* 1999, similar to that in Naidu *et al.* 2020), and have only the escape fraction as a free parameter.

N20 create and apply an empirical model to explore what objects carried out the bulk of ionization. Focusing specifically on ionizing photon escape fractions, their model ties the escape fraction to the star formation surface density (as motivated by recent samples of Lyman continuum leakers). Their model implies that rare very massive and UV bright galaxies ('oligarchs') account for the vast majority of the reionization budget.

When comparing these models purely on the reionization history they predict in Figure 4.1, N20 and Y20 follow a similar pattern: reionization starts very slowly and ramps up quickly after $z < 8$. N20 stands out among many reionization models

93

for how late and rapidly it occurs. F19 also stands out by presenting a reionization history that began earlier than $z > 11$ and evolved slowly and steadily. In the context of LAE observations with Roman, a universe that F19 describes will find many more LAEs at higher redshifts, as the IGM has more ionized gaps that allow the photons through. On the opposite side, N20 would predict very few LAEs at $z > 7$ and a mostly neutral IGM. Roman will hugely inform our understanding of reionization, with its large field-of-view and vast redshift coverage beyond the low-redshift universe where ground based observations have not found definitive constraints between these models.

4.2.3   Roman Grism Lyα Sensitivity

The final ingredient needed is Roman's expected sensitivity to Lyα across its broad redshift range. Here, we leverage some of the core results of Wold et al. (in prep). That work carries out detailed simulations of the LAE identification in Roman grism data, and quantifies the expected completeness of LAE samples across a range of luminosity and redshift. For this work, we use the results of their deepest simulation: 25 position angles at 10 kiloseconds each, for a total of 70 hours of exposure.

Wold et al. (in prep) test four bins for redshift completeness at z=7.5, 8.5, 9.5, and 10.5 for the Roman grism. They insert LAEs with known line flux and measured their recovered fraction. They have quantified what flux and Lyα line luminosity limits correspond to 50% completeness for various key redshifts. For $z = \{7.5, 8.5, 9.5, 10.5\}$, the limiting fluxes are $\{$1.4e-17, 8.3e-18, 7.5e-18, 6.9e-18$\}$ erg s^{-1} cm^{-2}, or Lyα line luminosities $\{$9.1e+42, 7.4e+42, 8.6e+42, 9.9e+42$\}$ erg s^{-1} respectively. The completeness functions approximate step functions, and we treat them as so in this work.

We show the redshift-dependent flux limits that we derive from Wold et al. (in prep) in Figure 4.2. Roman reaches significantly deeper limiting fluxes at the highest
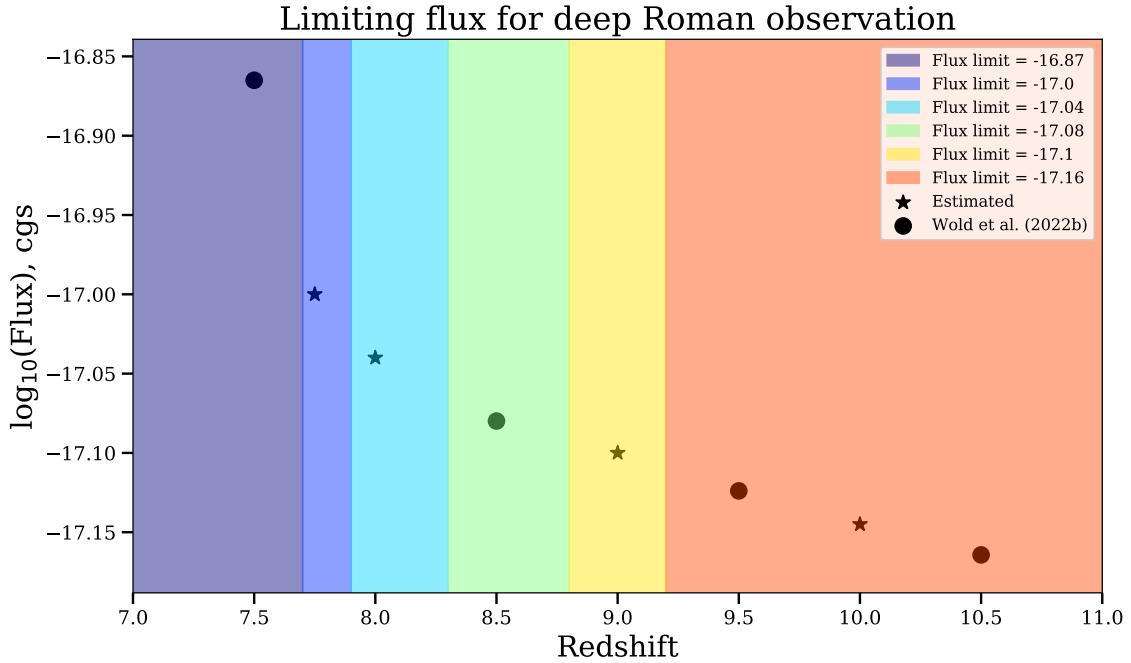
Figure 4.2: The redshift-dependent flux limits we use, as informed by the detailed Roman Lyα simulations of Wold et al. (in prep). The thresholds for applying each flux are indicated by the blocks of shaded color (shallowest at the lower redshifts in indigo, to the deepest high redshift observations in orange). Circles show the fluxes where 50% completeness of LAEs was directly simulated for a deep imaging; stars are our interpolations for in-between redshifts. We treat the completeness behavior as a step-function. The fluxes applied to each redshift-reionization scenario are listed in Table 4.1.
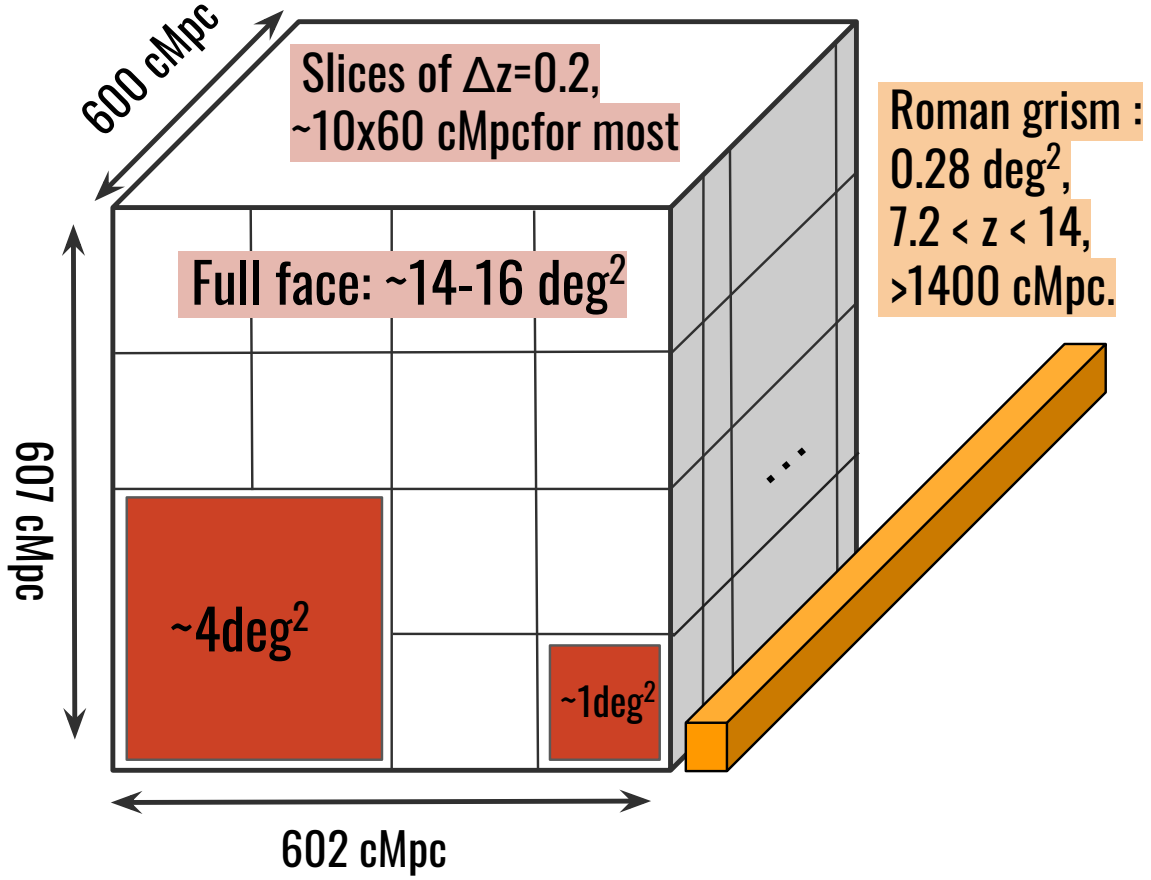
Figure 4.3: A schematic showing how we slice the J14 simulations to explore the constraints on reionization the Roman Space Telescope may find with the VPF. Depending on the redshift studied, the full face of the (601x607x600) cMpc^3 volume corresponds to approximately 14-16 deg^2. For our exploration of how the VPF evolves by redshift under different models for the reionization history of the universe, we test three survey areas with Roman: the full face ~ 16 deg^2, quarter-face ~ 4 deg^2, and sixteenth-face ~ 1 deg^2. We assume a survey covering $\Delta z = 0.2$, which yields between 7 to 14 even slices in the simulations. Table 4.1 detail each redshift-reionization scenario's slice depth and approximate LAE surface density.

redshifts for Lyα. The Roman grism sensitivity rises from its blue edge (1 μm, corresponding to $z = 7.2$ for Lyα) towards redder wavelengths. Balancing this effect against the rising luminosity distance, the Lyα luminosity limits are most sensitive near $z = 8.5$. Finally, we choose to apply a ceiling sensitivity of 1.4e-17 erg s^{-1} cm^{-2} for all redshifts under $z < 7.7$. For $z < 7.2$, the Roman prism can be used instead of the grism. Our assumption of a fixed flux limit at $z < 7.7$ is conservative, since the prism has higher throughput than the grism at $\sim 1\mu$m.

4.2.4   Projections for Roman VPF Measurements

We have described which discrete J14 ionization fraction simulations will likely be observed by Roman; when the models of reionization predict each; and the expected sensitivity of Roman to Lyα across its redshift range. We combine all these together to extract slices from the J14 simulations corresponding to the clustering analysis that Roman surveys will enable. In particular, we consider narrow slices (for a 2D VPF clustering analysis) and three different survey constructions: ~ 1 deg^2, ~ 4 deg^2, or ~ 16 deg^2.

Figure 4.3 illustrates how we slice up the J14 volumes. We assume focused studies of depth of $\Delta z \approx 0.2$, and create equally sized slices across the 600 cMpc volume depth. Depending on the redshift that a given ionization model predicts an ionization fraction, there can be between 14 slices $\times$ 43 cMpc/slice (e.g. F19 predicting $\langle x_i \rangle_v = 0.22$ at $z = 10.8$) to 7 slices $\times$ 86 cMpc/slice (e.g. F19 predicting $\langle x_i \rangle_v = 0.93$ at $z = 5.95$) in the z-direction. The columns under 'Slice$\times$Depth' in Table 4.1 detail this for each redshift-ionization scenario.

Now armed with mock $\Delta z \approx 0.2$ slices of the J14 simulations for each reionization model and redshift scenario, we apply a redshift-dependent flux limit (Figure 4.2, §4.2.3) to the LAEs in each slice. The simulated LAEs of J14 come with transmitted

97

Lyα luminosities. In order to apply the Roman-specific flux limit from Figure 4.2 to a given LAE, we translate the given flux limit to a luminosity cut based on the LAE's position within the slice. In each J14 full-face slice, we imagine the center Z-position is at the given redshift of focus, with $\Delta z \pm 0.1$ as the back and front edges of the slice. We associate the relative comoving position of a LAE in the slice to a luminosity cut, and apply given the flux limit for the central redshift in Figure 4.2 to generate a flux-limited samples for 2D clustering. We average the number of the LAEs that pass their redshift-position flux cut across all full-face slices, and divide by the area implied for 602×607 cMpc^2 at the central redshift. This yields the surface density "Σ, LAEs deg^{-2}" column in Table 4.1, ranging from a few dozen to a few hundred deg^{-2}. These are the LAE surface densities expected under the Wold et al. (in prep) sensitivity predictions and nuances of the J14 simulations across all our redshift-reionization scenarios.

Next, we consider the X and Y dimensions of our mock LAE samples. At the redshifts examined, the full-face of the (602×607) cMpc^2 J14 simulations cover between 13-16 deg^2. We split each slice into exact quarters (301×303.5 cMpc^2) or sixteenths (150.5×151.75 cMpc^2) to examine slices of approximately 4 deg^2 or 1 deg^2, respectively. This maximizes the number of completely independent mock LAE samples we are able to create, and covers a broad range of possible survey areas. Later, we will use all the independent $\Delta z = 0.2$ slices to explore the variability in the VPF as a function of survey area. Figure 4.3 displays our handling of the J14 simulations, with the additional illustration of what the default Roman grism will observe: an area of 0.281 deg^2 simultaneously covering $7.2 < z < 14$. Though we are artificially creating slices of $\Delta z = 0.2$ about specific redshifts, in truth Roman will be able to access LAEs across the bulk of reionization history. Table 4.1 summarizes the results of transforming each model into a redshift-reionization scenario for clustering

measurements.

4.2.5   The Void Probability Function for Reionization

The Void Probability Function (VPF) is a measurement of clustering that quantifies how likely a region of a given size is to be empty within the sample. Or, if one drops a certain number of circles of some R, how many are empty? It is the zero-eth moment of count-in-cells, focusing only on the cells with no galaxies, and therefore carries the signature of higher order volume-averaged clustering statistics (White 1979). Perez *et al.* (2021) applied an analysis of the hierarchical scaling between the VPF and higher order correlation functions (see also Conroy *et al.* 2005), comparing the VPF-derived volume averaged two-point correlation function and standard Landy and Szalay (1993) correlation function for simulated $3.1 < z < 6.6$ LAEs. They also derived theoretical descriptions for minimum and maximum distance scale guidelines for the VPF that we use in this work.

The VPF has also been leveraged specifically for reionization constraints. Kashikawa *et al.* (2006) and McQuinn *et al.* (2007) examined the ability of the VPF to constrain reionization for Subaru LAE observations at $z = 6.6$. More recently, Gangolli *et al.* (2021) explored the constraining power of the VPF and other clustering statistics for SILVERRUSH LAEs at $z = 5.7$. In Paper 1, we used the VPF to quantify what constraints on reionization the LAGER narrowband survey will be able to make at $z = 6.9$ with one, four, or eight fields' worth of imaging.

Table 4.1: Details for mock Roman observations given the Finkelstein *et al.* (2019, F19), Yung *et al.* (2020, Y20), and Naidu *et al.* (2020, N20) models for reionization, generated from to the Jensen *et al.* (2014, J14) volumes. For a given $\langle x_i \rangle_v$ simulation, we list: the redshift each model predicts the ionized fraction; the number and depth of slices corresponding to $\Delta z = 0.2$ depth; the flux limit applied at the redshift in $\log_{10}$ erg sec^{-1}; the corresponding limiting Lyα line luminosity in a slice; and the average surface density in deg^{-2} for one full-face slice.

$\langle x_i \rangle_v$	Slice Center z			Slices×Depth, cMpc			Lim. Flux, $\log_{10}$ erg s^{-1}			Lim. $\log_{10}L_{\text{Ly}\alpha}$			Σ, LAEs deg^{-2}		
Model:	F19	Y20	N20	F19	Y20	N20	F19	Y20	N20	F19	Y20	N20	F19	Y20	N20
0.22	10.8	8.75	7.6	14x43	12x50	10x60	-17.2	-17.08	-16.9	43.04	42.91	42.98	29	52	43
0.40	9.65	8.0	7.1	13x45	10x60	9x67	-17.2	-17.04	-16.9	42.92	42.88	42.91	66	103	85
0.485	9.1	7.75	6.9	12x50	10x60	8x75	-17.1	-17.0	-16.9	42.93	42.89	42.88	104	148	160
0.66	7.85	7.3	6.65	10x60	9x67	8x75	-17.0	-16.9	-16.9	42.88	42.97	42.85	162	133	211
0.93	5.95	6.85	6.3	7x86	8x75	8x75	-16.9	-16.9	-16.9	42.73	42.9	42.79	381	200	273

We measure the VPF using the k-nearest neighbors algorithm introduced in Banerjee and Abel (2020), which is notably faster than other methods. In a given slice, the transverse comoving positions of LAEs that passed the flux cut are normalized to cover e.g. $x = 0 - 301$, $y = 0 - 303.5$, and $z = 0 - 60$ cMpc for a ~ 4 deg^2 slice at $z = 9$. For each of several radii between 5 and 50 cMpc, we drop 100,000 random points in a slice to measure the VPF. We repeat this process 5 times to minimize sampling error, and use the average across the five samplings as the measured VPF of a given slice. When we later explore the VPF behavior of a given redshift-reionization scenario as a function of area, we will compare the mean and standard deviation of the VPF across all $\Delta z = 0.2$ slices of a particular area.

For a focused statistical analysis, in this work we will primarily discuss the VPF at $R \sim 12$ cMpc. $R \sim 12$ cMpc roughly corresponds to expected scale of ionized bubbles during reionization for the redshifts we examine. The smallest visible bubbles in the Lyα distribution are ones that will results in a Gunn and Peterson (1965) line center optical depth of $\tau \approx 1$, and which are therefore large enough to allow sufficient transmission of Lyα (see Rhoads and Malhotra 2001; Rhoads 2007). The characteristic radius of these bubbles comes out to 1.2 physical megaparsec, or:

$$R_{\text{bubble, cMpc}} = 1.2 \text{ pMpc} \times (1 + z). \qquad (4.1)$$

The VPF is a volume-averaged clustering statistic, which in practice means that general trends are consistent between nearby distance scales. In §4.3.3, however, we study the VPF across all radii for the models near $z = 7.75$.

4.3 Redshift Evolution of the VPF Across Reionization

We now examine the evolution of the VPF across reionization history that Roman will observe, and identify what constraints on reionization models the VPF may yield.
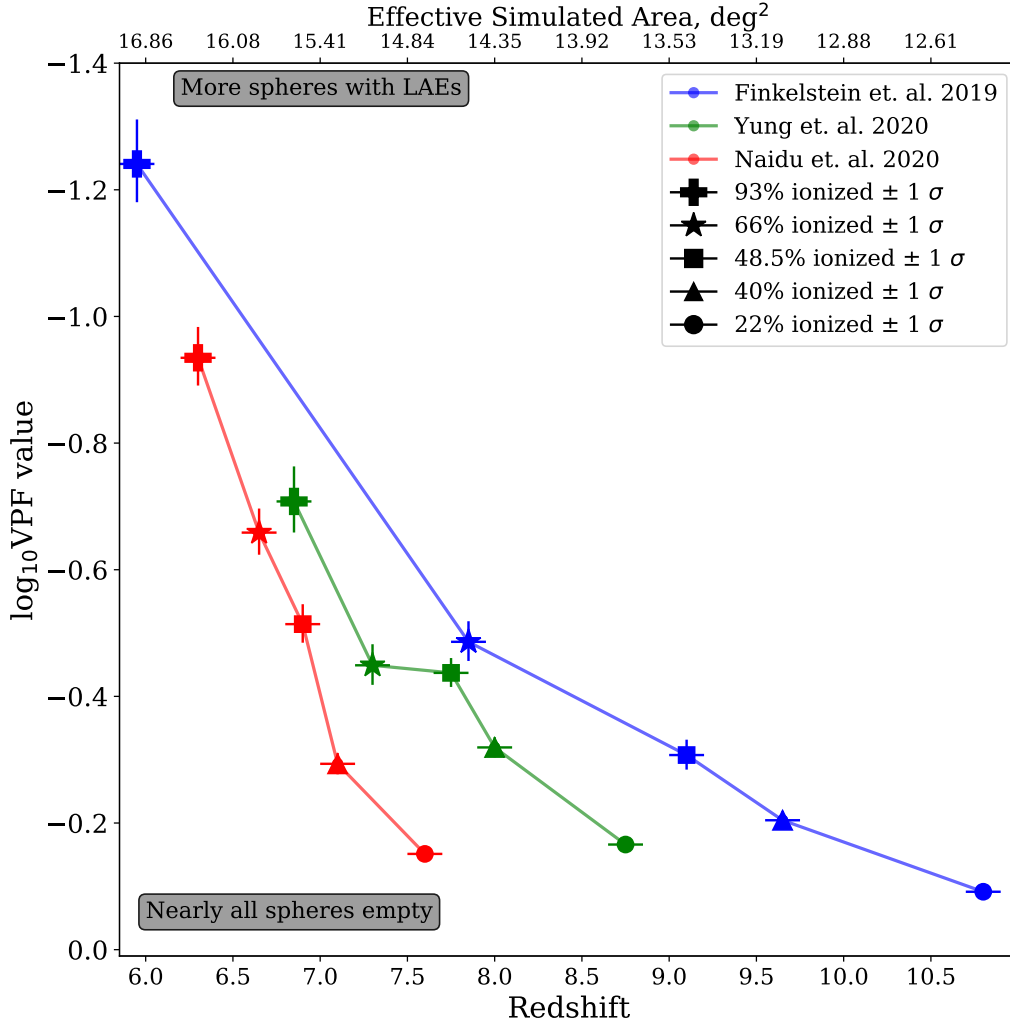
Figure 4.4: VPF(z) for a Roman field covering 14-16 deg^2 at 11.86 cMpc. The colors indicate the model of reionization: N20 (red), F19 (blue), and Y20 (green). The shapes of the points indicate which J14 volume was used for the VPF measurement: $\langle x_i \rangle_v = 93$ (plus sign, left-most), 66 (star), 48.5 (square), 40 (triangle), or 22 (circle, right-most) percent ionized. Error bars correspond to the 1σ standard deviation of the VPF across all full face $\sim 14 - 16$ deg^2 and $\Delta z = 0.2$ slices created for each simulation-reionization model pairing. The detailed VPF distributions can be studied in the Appendix.

We focus on three representative models of reionization: the very early and slow F19 model, the quick yet middle-of-the-road Y20 model, and the very late and very fast N20 model. The complete histograms for each model's redshift-reionization scenario VPF are shown at full scale in the Appendix B Figures 1 - 3. We ask: how well can Roman distinguish these reionization histories with the VPF using surveys of ~ 1 vs. ~ 4 vs. 13–16 deg^2?

4.3.1   Constraints from a 14-16 deg^2 Roman Survey

We first examine the results of the most ambitious survey we probe, for narrow slices encompassing 13-16 deg^2 of Roman spectroscopic observations. Figure 4.4 shows the VPF(z) at $R \sim 12$ cMpc for each redshift-reionization scenario, using several independent full-face (602x607) cMpc^2 slices from each simulation. The 3 representative models of reionization are indicated by different colors: N20 (red), F19 (blue), and Y20 (green). The shape of each symbol indicates which J14 simulation was sliced up for the VPF measurement: 22, 40, 48.5, 66, or 93 percent ionized (with markers circle, triangle, square, star, and plus, respectively). The error bars are the 1σ standard deviation across the 7-14 full-face slices' VPFs at the given radius (some of which are smaller than the symbol sizes).

These VPF(z) curves follow some trends: more neutral and higher-redshift slices have higher void probabilities (i.e. their $\log_{10}$(VPF) is closer to 0, meaning they have more empty dropped circles). This is not unsurprising, as lower density samples implicitly have more and larger voids. However, the pattern of the VPF also incorporates the predicted reionization histories, which Roman will be able to observe simultaneously across redshift history.

This largest explored survey area creates many opportunities for clear constraints on the timing and pace of reionization. First, we see the VPF is clearly different

between different ionization fractions at the same redshift (as seen in Paper 1, and further explored for $z = 7.75$ in §4.3.3). With this survey area, the VPF of all models' redshift-reionization scenarios are completely separated by at least 1.5σ (e.g. F19 and Y20 near $z = 7.8$), with most separated by $> 4\sigma$ from the nearest scenario of another model. For example, the Y20 ($\langle x_i \rangle_v$=0.93) and the N20 ($\langle x_i \rangle_v$=0.485) VPF measurements at $z = 7.8$ are distinguishable to 5σ, allowing Roman to decidedly constrain between a very late vs. somewhat early reionization. Past $z = 8$, where the Roman grism reaches peak sensitivity, the F19 ($\langle x_i \rangle_v$=0.66) and Y20 ($\langle x_i \rangle_v$=0.485) may be separated enough for distinguishing constraints of 3σ with a survey of $\sim$ 14 deg^2; therefore clarifying if reionization started very early or somewhat early. Additionally, the very late and fast N20 ($\langle x_i \rangle_v$=0.22) history can be distinguished from the earlier start of Y20 ($\langle x_i \rangle_v$=0.40, 0.485) or F19 ($\langle x_i \rangle_v$=0.66) to $> 8\sigma$ near $z = 7.75$. Finally, if the pacing and timing predicted by the F19 and Y20 VPF(z) hold near $z \sim 9$, the Roman may distinguish the two reionization histories to more than 4σ with the VPF.

Taken together, these results show that by combining any 2 to 3 redshifts, we are able to distinguish any pair of reionization histories with the VPF. More broadly, this VPF test will only be unable to distinguish two reionization models if they yield functionally identical reionization histories. Therefore, measurements of the Lyα VPF throughout the Roman grism can definitively identify which of the probed models best describes the reionization history of the universe.

4.3.2 Constraints from 4 and 1 deg^2 Roman Surveys

A LAE survey of 13-16 deg^2 with Roman would yield excellent $> 5\sigma$ constraints at several redshifts, and decisively rule in or our key features of reionization history. Might we reach similar constraints with smaller surveys? We repeat the process of
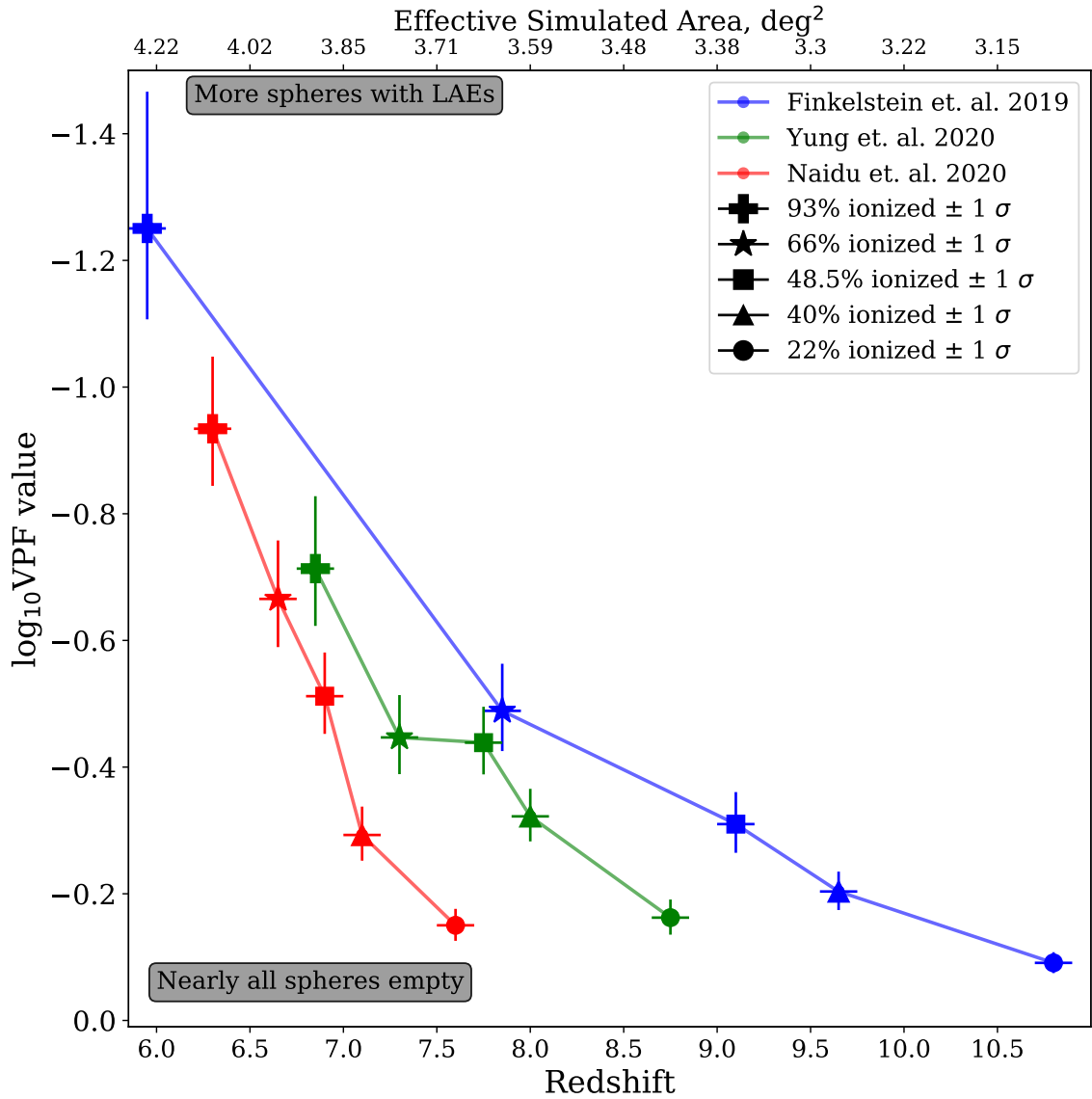
Figure 4.5: Like Figure 4.4, but instead for slices of ~ 4 deg^2 created for each simulation-reionization model pairing (at least 4×7 slices, each with a redshift depth of $\Delta z = 0.2$).

Figure 4.6: Like Figures 4.4 and 4.5, but instead for slices of approximately 1 deg^2 (one sixteenth of the face of the Jensen *et al.* 2014 simulations; at least 16×7 total slices). Error bars are the 1σ standard deviation across all slices' VPF.

Figure 4.5 using instead the ~ 4 deg^2 and ~ 1 deg^2 slices.

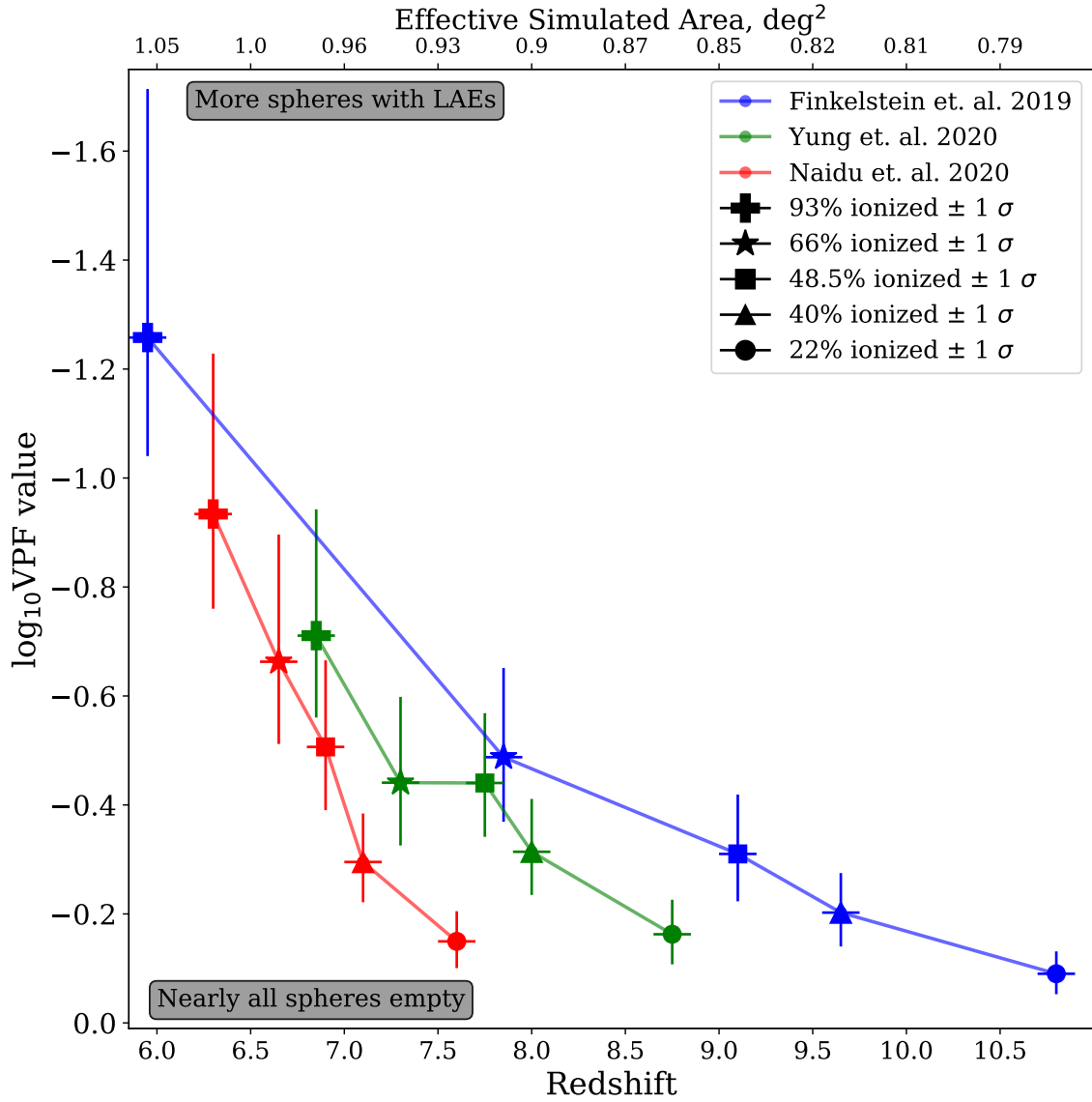The primary effect of the smaller area is decreased precision in the VPF from Poisson noise and larger variance across slices. However, we find the VPF across the ~ 4 deg^2 slices can still reach multiple constraints for the epoch and pace of reionization, to between 2.5-8σ across $6.8 < z < 9$. First, the Y20 ($\langle x_i \rangle_v$=0.93) and the N20 ($\langle x_i \rangle_v$=0.485) VPF measurements at $z \sim 6.8$ are distinguishable to about 2.5σ. This would identify a very late reionization at redshifts accessible with the Roman prism and ground-based surveys (consistent with our results in Paper 1 for LAGER). Additionally, near $z = 7.5 - 8$ the F19 ($\langle x_i \rangle_v$=0.66) and N20 ($\langle x_i \rangle_v$=0.22) scenarios could be distinguished to 3σ apart, independently constraining early vs. late and slow vs. fast reionization. Notably, more constraints for the timing of reionization are distinguishable to $6 - 8\sigma$ near $z = 7.75$ between the N20 ($\langle x_i \rangle_v$=0.22) and Y20 ($\langle x_i \rangle_v$=0.40, 0.485) models. However, we note these constraints are within the redshift range where the sensitivity of the grism changes rapidly, so LAE selection must be done carefully to leverage this survey area. Finally, the F19 ($\langle x_i \rangle_v$=0.485) and Y20 ($\langle x_i \rangle_v$=0.22) scenarios at $z \sim 9$ may be at least $3 - 4\sigma$ apart if the apparent patterns hold beyond where the discrete fractions of the J14 simulations reach.

Is it possible to make constraints with an even smaller survey? Figure 4.6 repeats our process for the ~ 1 deg^2 slices at $R = 11.86$ cMpc. The reduced area puts some of our earlier constraints out of reach, as fewer galaxies introduce more uncertainty to the VPF and makes fractions more difficult to distinguish. This scarcity of LAEs can still be a useful constraint: for example, an early investment of a few weeks of Roman observing time to cover 1 deg^2 may find fewer than 50 LAEs deg^{-2} at $z > 8$ (i.e. a low surface density consistent with the more neutral ionization fractions), which would provide some support for the later reionization models. However, there are still strong constraints the VPF of LAEs will yield for the timing of reionization. The

N20 ($\langle x_i \rangle_v$=0.22) and Y20 ($\langle x_i \rangle_v$=0.40, 0.485) histories near $z = 7.75$ are still distinguishable to $3 - 4\sigma$. Additionally, the F19 (near $\langle x_i \rangle_v$=0.485) and Y20 ($\langle x_i \rangle_v$=0.22) scenarios at $z \sim 9$ could perhaps be distinguished to more than 2σ apart, if the apparent trend of we see in the VPF(z) is true.

4.3.3   Example Focused Constraints near $7.6 < z < 7.9$

As we have explored, a constraint between very late vs. early reionization can be made near $7.6 < z < 7.9$ with all the survey areas we explore. Though we have so far focused on the behavior of the VPF at $R \sim 12$ cMpc, we measure the VPF between $5 < R < 50$ cMpc. How do the complete VPF curves behave for each of these distinct ionization predictions near $7.6 < z < 7.9$?

Figure 4.7 compares the complete VPF(R) measurements for the three models in this narrow redshift window: F19 predicting $\langle x_i \rangle_v$=0.66 near $z = 7.85$; Y20 predicting $\langle x_i \rangle_v$=0.485 near $z = 7.7$; and N20 predicting $\langle x_i \rangle_v$ <0.22 near $z = 7.6$. The dashed colored lines indicate the mean VPF across all the independent $\Delta z = 0.2$ slices of a given area. The colored shaded regions indicate the (log-space) errors corresponding to the 1σ standard deviation across the slices. The drop-off of the colored shading indicates where most slices start to find VPF=0. We also show as grey shading the maximum distance scale to measure the VPF for each survey area we probe (Perez *et al.*, 2021, given a precision to $\log_{10}$(VPF)> -1.5). We note that for the full-face 14-16 deg^2 surveys, the blue F19 and green Y20 1σ shadings begin to overlap only at $R > 15$ cMpc; and, that for the 4 deg^2 surveys, the mean Y20 VPF values are similar to the upper 1σ bound of the F20 shading at nearly all distances. Since N20 predicts such a late and swift reionization, we assume the ionization fraction will be no higher than 22% near $z = 7.75$, and that therefore the VPF measured for $\langle x_i \rangle_v$=0.22 near $z = 7.6$ is a good approximation. Though the Roman grism becomes more sensitive

past $z > 7.7$, we assume that this effect would be offset by a much more neutral IGM in the N20 model.

Much like Paper 1 showed for LAGER at $z = 6.9$, we find this single redshift can yield useful constraints on reionization even with smaller surveys. We find similar results for $z \sim 7.7$: increasing the area of the survey yields VPF measurements that more accurately reflect the 'true' VPF value of the sample with lower variance. We note that this work improves the VPF's precision by increasing the continuous covered area of a single imaging, rather than combining several independent smaller imagings of the sky as in Paper 1 with LAGER. With a ~ 1 deg^2 survey, N20 can be well-distinguished from the Y20/F19 scenarios ($> 3\sigma$), effectively ruling in or out a very late reionization history. Increasing to ~ 4 deg^2 improves the constraint notably ($> 6\sigma$) and also expands the distance scales that allow this constraint (to $R < 30$ cMpc). However, the Y20/F19 models cannot be distinguished until analyzing ~ 15 deg^2 ($\sim 1.5\sigma$ at $R < 15$ cMpc).

Finally, we remind readers that we have focused the scope of this analysis for two key reasons: first, we are limited to the ionized fractions of the existing J14 simulations; and second, we focus on the representative reionization models of F19, Y20, and N20, and where in cosmic history they predict each of the discrete fractions. Once actually observing with Roman, we will find completely independent measurements of the VPF at more redshifts than we are able to probe under this analysis. As we have done in Paper 1 for LAGER, the exercise of this sub-section can be done at any redshift to constrain which J14 ionization fraction simulation's clustering may best describe the VPF as measured in a deep Roman LAE survey. Therefore, the overall evolution of the VPF across $7.5 < z < 10.5$ will help confidently constrain the pacing of our universe's reionization history.
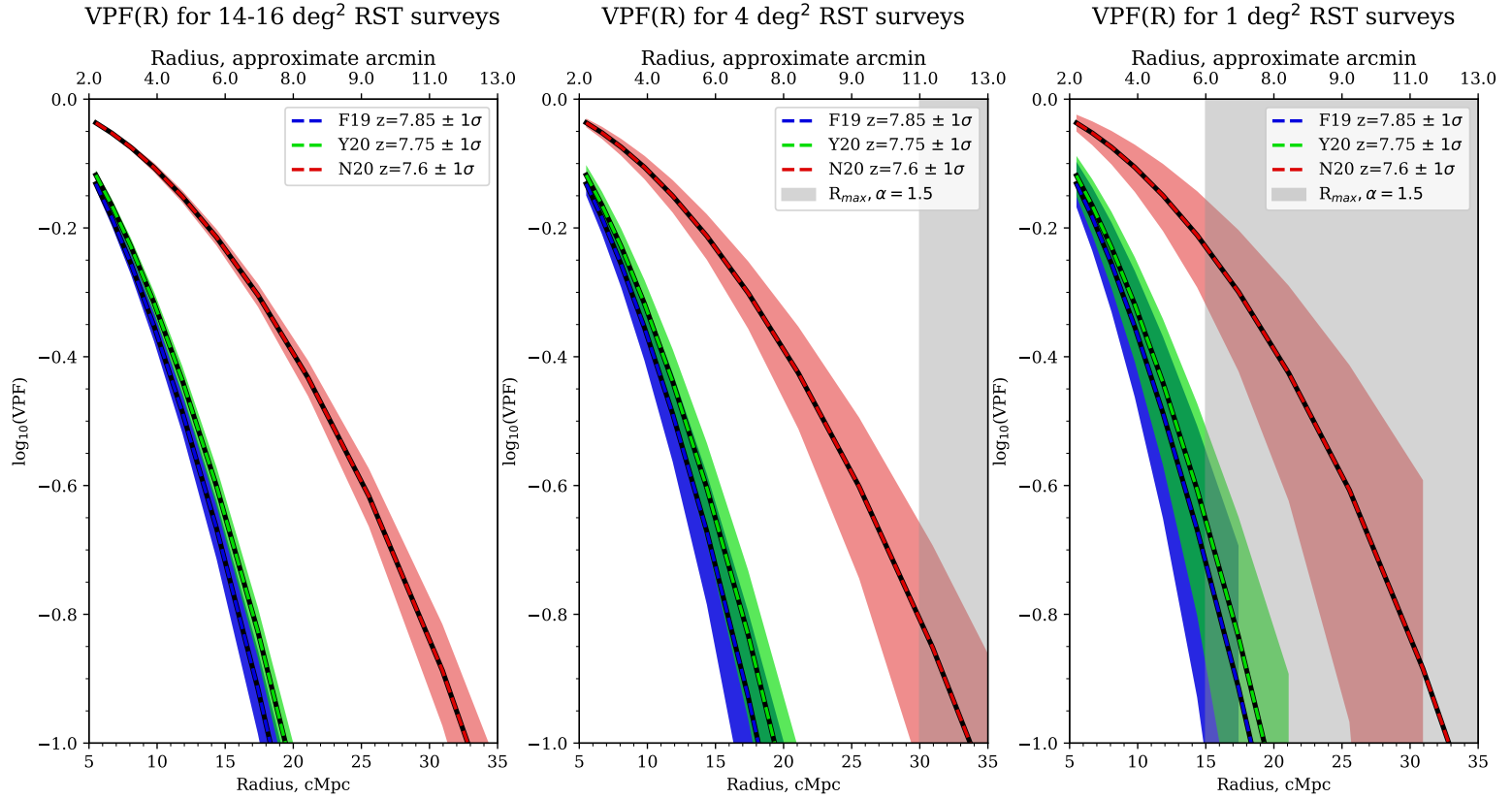
Figure 4.7: The complete VPF curves for the three ionization fractions–$\langle x_i \rangle_v$=0.22, 0.485, 0.66–predicted near $z \sim 7.75$ by the three models (N20, A20, F19 in red, green, blue respectively). The colored sharing are the 1σ standard deviation across all independent $\Delta z = 0.2$ slices of the given area. Grey shading indicates the approximate distance scales where the given survey area cannot measure the VPF (assuming $\alpha = 1.5$, and that all models exist at $z = 7.75$ for simplicity).

4.4 Conclusion

This work explores the constraints the Roman Space Telescope will find for the timing and pacing of reionization using the clustering of LAEs. We use the Void Probability Function (VPF), which asks how many randomly dropped spheres are empty, is tied to all higher order correlation functions, and is a simple clustering statistic to implement and guide survey construction. We focus on three representative models for reionization: Finkelstein *et al.* (2019, very early and slow reionization), Yung *et al.* (2020, early and quick reionization), and Naidu *et al.* (2020, late and fast reionization). We use the $(602 \times 607 \times 600)$ cMpc^3 simulations of LAEs through reionization of Jensen *et al.* (2014), at discrete ionized IGM fractions between $0.22 < \langle x_i \rangle_v < 0.93$. Informed by small and high-resolution simulations for Roman's grism responsiveness to Lyα (Wold et al. in prep), we create projections for LAEs as observed by Roman throughout reionization.

We mimic the three models' reionization histories by projecting each J14 simulation to the redshifts where each model predicts the given volume-averaged ionization fraction. We create mock Roman LAE surveys of $\Delta z = 0.2$ that cover: 14-16 deg^2 (the full face of the simulation), ~ 4 deg^2 (quarter face of the simulation), or ~ 1 deg^2 (one sixteenth of the simulation face). We measure the VPF and answer: what constraints on the pacing and timing of reionization might Roman find with the clustering of LAEs, as a function of survey area?

We find that a ~ 1 deg^2 survey can constrain between very late vs. early reionization near $z = 7.7$ to $> 3\sigma$, and perhaps also between very early and slow vs. quick and late reionization near $z \sim 9$ to $> 2\sigma$. By investing in 4 deg^2, the VPF of LAEs is able to distinguish: early vs. late reionization 3σ between $7.5 < z < 8$ and $z \sim 9$; and slow vs. fast reionization 3σ at $z \sim 8$. Additionally, if the ramp-up of the grism sensitivity

is considered, early and slow vs. fast and late reionization may be distinguished to 6σ or more near $z = 7.5 - 8$. Finally, we find a 13-16 deg^2 area would give VPF measurements that would essentially trace out a precise reionization history of the universe, and determine with great confidence ($> 4 - 5\sigma$) the most accurate model describing the timing and pacing of reionization.

Chapter 5

SUMMARY & OUTLOOK

I conclude this dissertation by summarizing our findings, and considering the future of the VPF and galaxy clustering. Detailed and complete summaries are given at the end of each Chapter.

In Chapter II, we applied the VPF to large simulations of Lyman-α emitters at $z = \{3.1, 4.5, 5.7, 6.6\}$ (Tilvi *et al.*, 2009), and examined the behavior of the VPF and the volume-averaged correlation function $\bar{\xi}$ under hierarchical scaling. We confirmed these LAEs are best fit by the negative binomial model (that also best describes this behavior in low-redshift galaxies), and probed how transforming the VPF into $\bar{\xi}$ (rather than how it is often done in reverse) informs these fits. Our results suggest that LAEs in the Tilvi *et al.* (2009) model may show large-scale higher clustering similar to that of low-redshift galaxies, and that larger simulations are needed to confirm this behavior. Additionally, we created general guidelines for when to measure the VPF, which we used in later Chapters to inform clustering surveys based off projected LAE surface densities and the amount of measured clustering.

In Chapter III, we shifted to assessing the core motivation of this dissertation: constraining reionization with the (VPF) clustering of LAEs. Using $(602\times607\times600)$ cMpc^3 large simulations of LAEs throughout reionization Jensen *et al.* (2014), we first focus specifically on informing future VPF analyses for the LAEs observed in the LAGER narrowband survey. We slice up and apply selections to the simulations to closely mimic LAGER LAE observations, yielding several hundred slices that can give us a sense of what a single LAGER DECam imaging may measure. We statistically probe what the VPF of LAGER might measure for one, four, or eight DECam

imagings of LAEs at $z = 6.9$: how well do each of these areas tell apart the ionized fractions of each volume? Our results suggest that even a single DECam field might be able to discriminate between mostly neutral or mostly ionized with the VPF. We also find that using four fields allows the distinction between 30, 50, and 95 percent (mass-averaged) ionized fraction, and the eight planned LAGER fields might even further distinguish between 73 and 95 percent ionized.

In Chapter IV, we expand our analysis to yield practical projections of how Roman may constrain reionization with the VPF of LAEs. We transform the Jensen *et al.* (2014) simulations to study different reionization histories: starting very early and occurring slowly (like Finkelstein *et al.* 2019), starting early and occurring somewhat quickly (like Yung *et al.* 2020), or starting very late and finishing very quickly (like Naidu *et al.* 2020). We apply selections to reflect how Roman will observe LAEs (Wold et al. 2022b, in prep), and create slices to reflect 2D clustering surveys of $\Delta z = 0.2$ and approximately 1, 4, or 16 deg^2. We identify that a 1 deg^2 survey will yield crucial initial constraints on an early vs. very late reionization to $> 3\sigma$ at $z \sim 7.7$, and also perhaps distinguish very early and slow vs. quick and late reionization to $> 2\sigma$ near $z \sim 9$. Investing in larger areas expands the accessible constraints and their robustness. We project a 4 deg^2 survey will be able to distinguish early vs. very late and slow vs. fast reionization to at least 3σ with several separate LAE samples. Finally, we show an ambitious 13-16 deg^2 survey will yield numerous robust $> 5\sigma$ constraints on the epoch and timing of reionization, and create a cohesive narrative for the EoR. We note these constraints are limited by the discrete ionization fractions we are able to study, the redshifts we are able to assess, and the complexities and caveats of all the models for reionization we compile.

As one may have noticed in the literature referenced throughout, the VPF is in a bit of a renaissance. We are not the only ones actively applying the VPF to reioniza-

tion in the last decade and a half (Gangolli *et al.*, 2021); folks are using it for better constraints on halo occupation (Walsh and Tinker, 2019); and recent computational innovations have featured its calculation (e.g. Banerjee and Abel 2020's kNN method, and Sinha and Garrison 2020 grouping all CiC calculations under the VPF). It would not be inaccurate to say I am happily attempting to lead this renaissance–as we move into the era of big data and precision cosmology, we should use every tool at our disposal, especially when it is as simple to implement and quietly profound as the VPF.

I will finish here by discussing what still needs to occur to maximize the knowledge we gain about the EoR with the VPF. Clearly, measuring the VPF of the observed LAGER LAEs is a goal of the near future. However, we also need robust measurements of the VPF of observed LAEs at lower redshifts. Getting a large enough sample of LAEs in a completely ionized IGM (but still high redshift enough to minimize inherent halo evolution) will be a crucial baseline for confirming the amount of clustering amplified by reionization. We also need larger areas and more LAEs. As alluded to in Chapters III and IV and their cited literature, another crucial constraint for the *process* of reionization requires more difficult galaxy selections. If we are able to create observed LAE samples at different redshifts/ionization fractions that are large enough to down-sample to a single number density, we will be able to quantify the additional clustering caused by inside-out reionization. This would be an exciting and unique constraint, that is finally within reach as Roman nears launch and operations.

REFERENCES

Adelberger, K. L., C. C. Steidel, M. Giavalisco, M. Dickinson, M. Pettini and M. Kellogg, "A Counts-in-Cells Analysis Of Lyman-break Galaxies At Redshift Z ~3", ApJ, **505**, 1, 18–24 (1998).

Ahmad, F., W. C. Saslaw and N. I. Bhat, "Statistical Mechanics of the Cosmological Many-Body Problem", ApJ, **571**, 2, 576–584 (2002).

Ajiki, M., Y. Taniguchi, S. S. Fujita, Y. Shioya, T. Nagao, T. Murayama, S. Yamada, K. Umeda and Y. Komiyama, "A Subaru Search for Lyα Emitters at Redshift 5.7", AJ, **126**, 5, 2091–2107 (2003).

Alimi, J.-M., A. Blanchard and R. Schaeffer, "Nonlinear Clustering in the CfA Redshift Survey", ApJL, **349**, L5 (1990).

Andrew, K., D. Barnaby and L. Taylor, "Analysis of Cosmological Generalized Reduced Void Probability Functions Constrained by Observations and Numerical Simulations", ArXiv e-prints (2013).

Astropy Collaboration, A. M. Price-Whelan, B. M. Sipőcz, H. M. Günther, P. L. Lim, S. M. Crawford, S. Conseil, D. L. Shupe, M. W. Craig, N. Dencheva, A. Ginsburg, J. T. Vand erPlas, L. D. Bradley, D. Pérez-Suárez, M. de Val-Borro, T. L. Aldcroft, K. L. Cruz, T. P. Robitaille, E. J. Tollerud, C. Ardelean, T. Babej, Y. P. Bach, M. Bachetti, A. V. Bakanov, S. P. Bamford, G. Barentsen, P. Barmby, A. Baumbach, K. L. Berry, F. Biscani, M. Boquien, K. A. Bostroem, L. G. Bouma, G. B. Brammer, E. M. Bray, H. Breytenbach, H. Buddelmeijer, D. J. Burke, G. Calderone, J. L. Cano Rodríguez, M. Cara, J. V. M. Cardoso, S. Cheedella, Y. Copin, L. Corrales, D. Crichton, D. D'Avella, C. Deil, É. Depagne, J. P. Dietrich, A. Donath, M. Droettboom, N. Earl, T. Erben, S. Fabbro, L. A. Ferreira, T. Finethy, R. T. Fox, L. H. Garrison, S. L. J. Gibbons, D. A. Goldstein, R. Gommers, J. P. Greco, P. Greenfield, A. M. Groener, F. Grollier, A. Hagen, P. Hirst, D. Homeier, A. J. Horton, G. Hosseinzadeh, L. Hu, J. S. Hunkeler, Ž. Ivezić, A. Jain, T. Jenness, G. Kanarek, S. Kendrew, N. S. Kern, W. E. Kerzendorf, A. Khvalko, J. King, D. Kirkby, A. M. Kulkarni, A. Kumar, A. Lee, D. Lenz, S. P. Littlefair, Z. Ma, D. M. Macleod, M. Mastropietro, C. McCully, S. Montagnac, B. M. Morris, M. Mueller, S. J. Mumford, D. Muna, N. A. Murphy, S. Nelson, G. H. Nguyen, J. P. Ninan, M. Nöthe, S. Ogaz, S. Oh, J. K. Parejko, N. Parley, S. Pascual, R. Patil, A. A. Patil, A. L. Plunkett, J. X. Prochaska, T. Rastogi, V. Reddy Janga, J. Sabater, P. Sakurikar, M. Seifert, L. E. Sherbert, H. Sherwood-Taylor, A. Y. Shih, J. Sick, M. T. Silbiger, S. Singanamalla, L. P. Singer, P. H. Sladen, K. A. Sooley, S. Sornarajah, O. Streicher, P. Teuben, S. W. Thomas, G. R. Tremblay, J. E. H. Turner, V. Terrón, M. H. van Kerkwijk, A. de la Vega, L. L. Watkins, B. A. Weaver, J. B. Whitmore, J. Woillez, V. Zabalza and Astropy Contributors, "The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package", AJ, **156**, 3, 123 (2018).

Astropy Collaboration, T. P. Robitaille, E. J. Tollerud, P. Greenfield, M. Droett-boom, E. Bray, T. Aldcroft, M. Davis, A. Ginsburg, A. M. Price-Whelan, W. E. Kerzendorf, A. Conley, N. Crighton, K. Barbary, D. Muna, H. Ferguson, F. Grollier, M. M. Parikh, P. H. Nair, H. M. Unther, C. Deil, J. Woillez, S. Conseil, R. Kramer, J. E. H. Turner, L. Singer, R. Fox, B. A. Weaver, V. Zabalza, Z. I. Edwards, K. Azalee Bostroem, D. J. Burke, A. R. Casey, S. M. Crawford, N. Dencheva, J. Ely, T. Jenness, K. Labrie, P. L. Lim, F. Pierfederici, A. Pontzen, A. Ptak, B. Refsdal, M. Servillat and O. Streicher, "Astropy: A community Python package for astronomy", A&A, **558**, A33 (2013).

Atek, H., D. Kunth, D. Schaerer, J. M. Mas-Hesse, M. Hayes, G. Östlin and J.-P. Kneib, "Influence of physical galaxy properties on Lyα escape in star-forming galaxies", A&A, **561**, A89 (2014).

Banerjee, A. and T. Abel, "Nearest Neighbor distributions: new statistical measures for cosmological clustering", arXiv e-prints p. arXiv:2007.13342 (2020).

Baugh, C. M., D. J. Croton, E. Gaztañaga, P. Norberg, M. Colless, I. K. Baldry, J. Bland -Hawthorn, T. Bridges, R. Cannon, S. Cole, C. Collins, W. Couch, G. Dalton, R. De Propris, S. P. Driver, G. Efstathiou, R. S. Ellis, C. S. Frenk, K. Glazebrook, C. Jackson, O. Lahav, I. Lewis, S. Lumsden, S. Maddox, D. Madgwick, J. A. Peacock, B. A. Peterson, W. Sutherland, K. Taylor and 2dFGRS Team, "The 2dF Galaxy Redshift Survey: hierarchical galaxy clustering", MNRAS, **351**, 2, L44–L48 (2004).

Bayer, A. E., F. Villaescusa-Navarro, E. Massara, J. Liu, D. N. Spergel, L. Verde, B. D. Wandelt, M. Viel and S. Ho, "Detecting Neutrino Mass by Combining Matter Clustering, Halos, and Voids", ApJ, **919**, 1, 24 (2021).

Bel, J., E. Branchini, C. Di Porto, O. Cucciati, B. R. Granett, A. Iovino, S. de la Torre, C. Marinoni, L. Guzzo, L. Moscardini, A. Cappi, U. Abbas, C. Adami, S. Arnouts, M. Bolzonella, D. Bottini, J. Coupon, I. Davidzon, G. De Lucia, A. Fritz, P. Franzetti, M. Fumana, B. Garilli, O. Ilbert, J. Krywult, V. Le Brun, O. Le Fèvre, D. Maccagni, K. Małek, F. Marulli, H. J. McCracken, L. Paioro, M. Polletta, A. Pollo, H. Schlagenhaufer, M. Scodeggio, L. A. M. Tasca, R. Tojeiro, D. Vergani, A. Zanichelli, A. Burden, A. Marchetti, Y. Mellier, R. C. Nichol, J. A. Peacock, W. J. Percival, S. Phleps and M. Wolk, "The VIMOS Public Extragalactic Redshift Survey (VIPERS). On the recovery of the count-in-cell probability distribution function", A&A, **588**, A51 (2016).

Beltz-Mohrmann, G. D., A. A. Berlind and A. O. Szewciw, "Testing the accuracy of halo occupation distribution modelling using hydrodynamic simulations", MNRAS, **491**, 4, 5771–5788 (2020).

Benson, A. J., "Galaxy formation theory", Phys. Rep., **495**, 33–86 (2010).

Berlind, A. A. and D. H. Weinberg, "The Halo Occupation Distribution: Toward an Empirical Determination of the Relation between Galaxies and Mass", ApJ, **575**, 587–616 (2002).

Bernardeau, F., S. Colombi, E. Gaztañaga and R. Scoccimarro, "Large-scale structure of the Universe and cosmological perturbation theory", Phys. Rep., **367**, 1–248 (2002).

Betancort-Rijo, J., "Generalized negative binomial distributions", Journal of Statistical Physics **98**, 3, 917–933, URL `https://doi.org/10.1023/A:1018691912596` (2000).

Bielby, R. M., P. Tummuangpak, T. Shanks, H. Francke, N. H. M. Crighton, E. Bañados, J. González-López, L. Infante and A. Orsi, "The VLT LBG redshift survey - V. Characterizing the $z = 3.1$ Lyman α emitter population", MNRAS, **456**, 4061–4080 (2016).

Carruthers, P. and M. Duong-van, "A connection between galaxy probabilities in zwicky clusters counting distributions in particle physics and quantum optics", Physics Letters B **131**, 1-3, 116–120 (1983).

Coil, A. L., *The Large-Scale Structure of the Universe*, vol. 6, p. 384 (Springer Science+Business Media Dordrecht, 2013).

Conroy, C., A. L. Coil, M. White, J. A. Newman, R. Yan, M. C. Cooper, B. F. Gerke, M. Davis and D. C. Koo, "The DEEP2 Galaxy Redshift Survey: The Evolution of Void Statistics from z ˜ 1 to z ˜ 0", ApJ, **635**, 990–1005 (2005).

Cooper, M. C., J. A. Newman, A. L. Coil, D. J. Croton, B. F. Gerke, R. Yan, M. Davis, S. M. Faber, P. Guhathakurta, D. C. Koo, B. J. Weiner and C. N. A. Willmer, "The DEEP2 galaxy redshift survey: evolution of the colour-density relation at 0.4 ¡ z ¡ 1.35", MNRAS, **376**, 4, 1445–1459 (2007).

Crocce, M., S. Pueblas and R. Scoccimarro, "Transients from initial conditions in cosmological simulations", MNRAS, **373**, 369–381 (2006).

Croton, D. J., M. Colless, E. Gaztañaga, C. M. Baugh, P. Norberg, I. K. Baldry, J. Bland-Hawthorn, T. Bridges, R. Cannon, S. Cole, C. Collins, W. Couch, G. Dalton, R. de Propris, S. P. Driver, G. Efstathiou, R. S. Ellis, C. S. Frenk, K. Glazebrook, C. Jackson, O. Lahav, I. Lewis, S. Lumsden, S. Maddox, D. Madgwick, J. A. Peacock, B. A. Peterson, W. Sutherland and K. Taylor, "The 2dF Galaxy Redshift Survey: voids and hierarchical scaling models", MNRAS, **352**, 828–836 (2004a).

Croton, D. J., G. R. Farrar, P. Norberg, M. Colless, J. A. Peacock, I. K. Baldry, C. M. Baugh, J. Bland-Hawthorn, T. Bridges, R. Cannon, S. Cole, C. Collins, W. Couch, G. Dalton, R. De Propris, S. P. Driver, G. Efstathiou, R. S. Ellis, C. S. Frenk, K. Glazebrook, C. Jackson, O. Lahav, I. Lewis, S. Lumsden, S. Maddox, D. Madgwick, B. A. Peterson, W. Sutherland and K. Taylor, "The 2dF Galaxy Redshift Survey: luminosity functions by density environment and galaxy type", MNRAS, **356**, 3, 1155–1167 (2005).

Croton, D. J., E. Gaztañaga, C. M. Baugh, P. Norberg, M. Colless, I. K. Baldry, J. Bland -Hawthorn, T. Bridges, R. Cannon, S. Cole, C. Collins, W. Couch, G. Dalton, R. De Propris, S. P. Driver, G. Efstathiou, R. S. Ellis, C. S. Frenk, K. Glazebrook, C. Jackson, O. Lahav, I. Lewis, S. Lumsden, S. Maddox, D. Madgwick, J. A.

Peacock, B. A. Peterson, W. Sutherland and K. Taylor, "The 2dF Galaxy Redshift Survey: higher-order galaxy correlation functions", MNRAS, **352**, 4, 1232–1244 (2004b).

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

Dayal, P., A. Ferrara and A. Saro, "The cool side of Lyman alpha emitters", MNRAS, **402**, 3, 1449–1457 (2010).

Dayal, P., A. Ferrara, A. Saro, R. Salvaterra, S. Borgani and L. Tornatore, "Lyman alpha emitter evolution in the reionization epoch", MNRAS, **400**, 4, 2000–2011 (2009).

Dayal, P., A. Maselli and A. Ferrara, "The visibility of Lyman α emitters during reionization", MNRAS, **410**, 2, 830–843 (2011).

Dijkstra, M., J. S. B. Wyithe and Z. Haiman, "Luminosity functions of Lyα emitting galaxies and cosmic reionization of hydrogen", MNRAS, **379**, 1, 253–259 (2007).

Dunkley, J., E. Komatsu, M. R. Nolta, D. N. Spergel, D. Larson, G. Hinshaw, L. Page, C. L. Bennett, B. Gold, N. Jarosik, J. L. Weiland, M. Halpern, R. S. Hill, A. Kogut, M. Limon, S. S. Meyer, G. S. Tucker, E. Wollack and E. L. Wright, "Five-Year Wilkinson Microwave Anisotropy Probe Observations: Likelihoods and Parameters from the WMAP Data", ApJS, **180**, 2, 306–329 (2009).

Elizalde, E. and E. Gaztanaga, "Void probability as a function of the void's shape and scale-invariant models", MNRAS, **254**, 247–256 (1992).

Endsley, R., D. P. Stark, S. Charlot, J. Chevallard, B. Robertson, R. J. Bouwens and M. Stefanon, "MMT spectroscopy of lyman-alpha at z7: Evidence for accelerated reionization around massive galaxies", MNRAS, (2021).

Fan, X., M. A. Strauss, R. H. Becker, R. L. White, J. E. Gunn, G. R. Knapp, G. T. Richards, D. P. Schneider, J. Brinkmann and M. Fukugita, "Constraining the Evolution of the Ionizing Background and the Epoch of Reionization with z~6 Quasars. II. A Sample of 19 Quasars", AJ, **132**, 1, 117–136 (2006).

Finkelstein, S. L., A. D'Aloisio, J.-P. Paardekooper, J. Ryan, Russell, P. Behroozi, K. Finlator, R. Livermore, P. R. Upton Sanderbeck, C. Dalla Vecchia and S. Khochfar, "Conditions for Reionizing the Universe with a Low Galaxy Ionizing Photon Escape Fraction", ApJ, **879**, 1, 36 (2019).

Finkelstein, S. L., J. E. Rhoads, S. Malhotra, N. Grogin and J. Wang, "Effects of Dust Geometry in Lyα Galaxies at z = 4.4", ApJ, **678**, 2, 655–668 (2008).

Finkelstein, S. L., J. E. Rhoads, S. Malhotra, N. Pirzkal and J. Wang, "The Ages and Masses of Lyα Galaxies at z ~4.5", ApJ, **660**, 2, 1023–1029 (2007).

Friedrich, M. M., G. Mellema, M. A. Alvarez, P. R. Shapiro and I. T. Iliev, "Topology and sizes of H II regions during cosmic reionization", MNRAS, **413**, 2, 1353–1372 (2011).

Fry, J. N., "Cosmological density fluctuations and large-scale structure From N-point correlation functions to the probability distribution", ApJ, **289**, 10–17 (1985).

Fry, J. N., "Statistics of Voids in Hierarchical Universes", ApJ, **306**, 358 (1986).

Fry, J. N., "VI. Voids and Scaling in Cosmology", Publications of the Astronomical Society of the Pacific **100**, 1336 (1988).

Fry, J. N. and S. Colombi, "Void statistics and hierarchical scaling in the halo model", MNRAS, **433**, 581–590 (2013).

Fry, J. N., R. Giovanelli, M. P. Haynes, A. L. Melott and R. J. Scherrer, "Void Statistics, Scaling, and the Origins of Large-Scale Structure", ApJ, **340**, 11 (1989).

Furlanetto, S. R., L. Hernquist and M. Zaldarriaga, "Constraining the topology of reionization through Lyα absorption", MNRAS, **354**, 3, 695–707 (2004).

Furlanetto, S. R., M. Zaldarriaga and L. Hernquist, "The effects of reionization on Lyα galaxy surveys", MNRAS, **365**, 3, 1012–1020 (2006).

Gangolli, N., A. D'Aloisio, F. Nasir and Z. Zheng, "Constraining reionization in progress at z = 5.7 with Lyman-α emitters: voids, peaks, and cosmic variance", MNRAS, **501**, 4, 5294–5308 (2021).

Gawiser, E., H. Francke, K. Lai, K. Schawinski, C. Gronwall, R. Ciardullo, R. Quadri, A. Orsi, L. F. Barrientos, G. A. Blanc, G. Fazio, J. J. Feldmeier, J.-s. Huang, L. Infante, P. Lira, N. Padilla, E. N. Taylor, E. Treister, C. M. Urry, P. G. van Dokkum and S. N. Virani, "Lyα-Emitting Galaxies at z = 3.1: L* Progenitors Experiencing Rapid Star Formation", ApJ, **671**, 278–284 (2007).

Gaztanaga, E. and J. Yokoyama, "Probing the statistics of primordial fluctuations and their evolution", ApJ, **403**, 450–465 (1993).

Gronwall, C., R. Ciardullo, T. Hickey, E. Gawiser, J. J. Feldmeier, P. G. van Dokkum, C. M. Urry, D. Herrera, B. D. Lehmer, L. Infante, A. Orsi, D. Marchesini, G. A. Blanc, H. Francke, P. Lira and E. Treister, "Lyα Emission-Line Galaxies at z = 3.1 in the Extended Chandra Deep Field-South", ApJ, **667**, 79–91 (2007).

Guaita, L., E. Gawiser, N. Padilla, H. Francke, N. A. Bond, C. Gronwall, R. Ciardullo, J. J. Feldmeier, S. Sinawa, G. A. Blanc and S. Virani, "Lyα-emitting Galaxies at z = 2.1 in ECDF-S: Building Blocks of Typical Present-day Galaxies?", ApJ, **714**, 255–269 (2010).

Gunn, J. E. and B. A. Peterson, "On the Density of Neutral Hydrogen in Intergalactic Space.", ApJ, **142**, 1633–1636 (1965).

Habouzit, M., A. Pisani, A. Goulding, Y. Dubois, R. S. Somerville and J. E. Greene, "Properties of simulated galaxies and supermassive black holes in cosmic voids", MNRAS, **493**, 1, 899–921 (2020).

Haiman, Z. and R. Cen, "Constraining Reionization with the Evolution of the Luminosity Function of Lyα Emitting Galaxies", ApJ, **623**, 2, 627–631 (2005).

Haiman, Z. and M. Spaans, "Models for Dusty Lyα Emitters at High Redshift", ApJ, **518**, 1, 138–144 (1999).

Hamaus, N., A. Pisani, P. M. Sutter, G. Lavaux, S. Escoffier, B. D. Wandelt and J. Weller, "Constraints on Cosmology and Gravity from the Dynamics of Voids", Phys. Rev. Lett., **117**, 9, 091302 (2016).

Hamilton, A. J. S., "On Hierarchical Solutions to the BBGKY Hierarchy", ApJ, **332**, 67 (1988).

Hao, C.-N., J.-S. Huang, X. Xia, X. Zheng, C. Jiang and C. Li, "A Deep Lyα Survey in ECDF-S and COSMOS. I. General Properties of Lyα Emitters at z $\sim$ 2", ApJ, **864**, 145 (2018).

Harikane, Y., M. Ouchi, T. Shibuya, T. Kojima, H. Zhang, R. Itoh, Y. Ono, R. Higuchi, A. K. Inoue, J. Chevallard, P. L. Capak, T. Nagao, M. Onodera, A. L. Faisst, C. L. Martin, M. Rauch, G. A. Bruzual, S. Charlot, I. Davidzon, S. Fujimoto, M. Hilmi, O. Ilbert, C.-H. Lee, Y. Matsuoka, J. D. Silverman and S. Toft, "SILVERRUSH. V. Census of Lyα, [O III] λ5007, Hα, and [C II] 158 μm Line Emission with 1000 LAEs at z = 4.9-7.0 Revealed with Subaru/HSC", ApJ, **859**, 2, 84 (2018).

Harish, S., I. G. B. Wold, S. Malhotra, J. Rhoads, W. Hu, J. Wang, Z.-y. Zheng, L. F. Barrientos, J. González-López, L. A. Perez, A. A. Khostovan, L. Infante, C. Jiang, C. Moya-Sierralta, J. Pharo, F. Valdes and H. Yang, "New spectroscopic confirmations of Lyman-α emitters at z $\sim$ 7 from the LAGER survey", arXiv e-prints p. arXiv:2111.01173 (2021).

Hayes, M., G. Östlin, F. Duval, A. Sandberg, L. Guaita, J. Melinder, A. Adamo, D. Schaerer, A. Verhamme, I. Orlitová, J. M. Mas-Hesse, J. M. Cannon, H. Atek, D. Kunth, P. Laursen, H. Otí-Floranes, S. Pardy, T. Rivera-Thorsen and E. C. Herenz, "The Lyman Alpha Reference Sample. II. Hubble Space Telescope Imaging Results, Integrated Properties, and Trends", ApJ, **782**, 1, 6 (2014).

Hinshaw, G., D. Larson, E. Komatsu, D. N. Spergel, C. L. Bennett, J. Dunkley, M. R. Nolta, M. Halpern, R. S. Hill, N. Odegard, L. Page, K. M. Smith, J. L. Weiland, B. Gold, N. Jarosik, A. Kogut, M. Limon, S. S. Meyer, G. S. Tucker, E. Wollack and E. L. Wright, "Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results", ApJS, **208**, 2, 19 (2013).

Hong, S., A. Dey, K.-S. Lee, Á. A. Orsi, K. Gebhardt, M. Vogelsberger, L. Hernquist, R. Xue, I. Jung, S. L. Finklestein, S. Tuttle and M. Boylan- Kolchin, "Statistics

of two-point correlation and network topology for Ly α emitters at z $\approx$ 2.67", MNRAS, **483**, 3950–3970 (2019).

Hu, E. M., L. L. Cowie, P. Capak, R. G. McMahon, T. Hayashino and Y. Komiyama, "The Luminosity Function of Lyα Emitters at Redshift z~5.7", AJ, **127**, 2, 563–575 (2004).

Hu, E. M., L. L. Cowie, R. G. McMahon, P. Capak, F. Iwamuro, J. P. Kneib, T. Maihara and K. Motohara, "A Redshift z=6.56 Galaxy behind the Cluster Abell 370", ApJL, **568**, 2, L75–L79 (2002).

Hu, W., J. Wang, L. Infante, J. E. Rhoads, Z.-Y. Zheng, H. Yang, S. Malhotra, L. F. Barrientos, C. Jiang, J. González-López, G. Prieto, L. A. Perez, P. Hibon, G. Galaz, A. Coughlin, S. Harish, X. Kong, W. Kang, A. A. Khostovan, J. Pharo, F. Valdes, I. Wold, A. R. Walker and X. Zheng, "A Lyman-α protocluster at redshift 6.9", Nature Astronomy (2021).

Hu, W., J. Wang, Z.-Y. Zheng, S. Malhotra, L. Infante, J. Rhoads, A. Gonzalez, A. R. Walker, L. Jiang, C. Jiang, P. Hibon, L. F. Barrientos, S. Finkelstein, G. Galaz, W. Kang, X. Kong, V. Tilvi, H. Yang and X. Zheng, "First Spectroscopic Confirmations of z $\sim$ 7.0 Lyα Emitting Galaxies in the LAGER Survey", ApJL, **845**, 2, L16 (2017).

Hu, W., J. Wang, Z.-Y. Zheng, S. Malhotra, J. E. Rhoads, L. Infante, L. F. Barrientos, H. Yang, C. Jiang, W. Kang, L. A. Perez, I. Wold, P. Hibon, L. Jiang, A. A. Khostovan, F. Valdes, A. R. Walker, G. Galaz, A. Coughlin, S. Harish, X. Kong, J. Pharo and X. Zheng, "The Lyα Luminosity Function and Cosmic Reionization at z $\sim$ 7.0: A Tale of Two LAGER Fields", ApJ, **886**, 2, 90 (2019).

Hurtado-Gil, L., V. J. Martínez, P. Arnalte- Mur, M.-J. Pons-Bordería, C. Pareja-Flores and S. Paredes, "The best fit for the observed galaxy counts-in-cell distribution function", A&A, **601**, A40 (2017).

Iliev, I. T., G. Mellema, P. R. Shapiro, U.-L. Pen, Y. Mao, J. Koda and K. Ahn, "Can 21-cm observations discriminate between high-mass and low-mass galaxies as reionization sources?", MNRAS, **423**, 3, 2222–2253 (2012).

Iliev, I. T., P. R. Shapiro, P. McDonald, G. Mellema and U.-L. Pen, "The effect of the intergalactic environment on the observability of Lyα emitters during reionization", MNRAS, **391**, 1, 63–83 (2008).

Inoue, A. K., K. Hasegawa, T. Ishiyama, H. Yajima, I. Shimizu, M. Umemura, A. Konno, Y. Harikane, T. Shibuya, M. Ouchi, K. Shimasaku, Y. Ono, H. Kusakabe, R. Higuchi and C.-H. Lee, "SILVERRUSH. VI. A simulation of Lyα emitters in the reionization epoch and a comparison with Subaru Hyper Suprime-Cam survey early data", PASJ, **70**, 3, 55 (2018).

Iye, M., K. Ota, N. Kashikawa, H. Furusawa, T. Hashimoto, T. Hattori, Y. Matsuda, T. Morokuma, M. Ouchi and K. Shimasaku, "A galaxy at a redshift z = 6.96", Nature, **443**, 7108, 186–188 (2006).

Jensen, H., M. Hayes, I. T. Iliev, P. Laursen, G. Mellema and E. Zackrisson, "Studying reionization with the next generation of Lyα emitter surveys", MNRAS, **444**, 3, 2114–2127 (2014).

Jensen, H., P. Laursen, G. Mellema, I. T. Iliev, J. Sommer-Larsen and P. R. Shapiro, "On the use of Lyα emitters as probes of reionization", MNRAS, **428**, 2, 1366–1381 (2013).

Jiang, L., E. Egami, M. Mechtley, X. Fan, S. H. Cohen, R. A. Windhorst, R. Davé, K. Finlator, N. Kashikawa, M. Ouchi and K. Shimasaku, "Physical Properties of Spectroscopically Confirmed Galaxies at z ¿= 6. I. Basic Characteristics of the Rest-frame UV Continuum and Lyα Emission", ApJ, **772**, 2, 99 (2013).

Jones, E., T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python", URL `http://www.scipy.org/` (2001–).

Kaiser, N., "Clustering in real space and in redshift space", MNRAS, **227**, 1–21 (1987).

Kakiichi, K., M. Dijkstra, B. Ciardi and L. Graziani, "Lyα-emitting galaxies as a probe of reionization: large-scale bubble morphology and small-scale absorbers", MNRAS, **463**, 4, 4019–4040 (2016).

Kashikawa, N., K. Shimasaku, M. A. Malkan, M. Doi, Y. Matsuda, M. Ouchi, Y. Taniguchi, C. Ly, T. Nagao, M. Iye, K. Motohara, T. Murayama, K. Muro-zono, K. Nariai, K. Ohta, S. Okamura, T. Sasaki, Y. Shioya and M. Umemura, "The End of the Reionization Epoch Probed by Lyα Emitters at z = 6.5 in the Subaru Deep Field", ApJ, **648**, 7–22 (2006).

Kashikawa, N., K. Shimasaku, Y. Matsuda, E. Egami, L. Jiang, T. Nagao, M. Ouchi, M. A. Malkan, T. Hattori, K. Ota, Y. Taniguchi, S. Okamura, C. Ly, M. Iye, H. Furusawa, Y. Shioya, T. Shibuya, Y. Ishizaki and J. Toshikawa, "Completing the Census of Lyα Emitters at the Reionization Epoch", ApJ, **734**, 2, 119 (2011).

Khostovan, A. A., D. Sobral, B. Mobasher, P. N. Best, I. Smail, J. Matthee, B. Darvish, H. Nayyeri, S. Hemmati and J. P. Stott, "The clustering of H β + [O III] and [O II] emitters since z 5: dependencies with line luminosity and stellar mass", MNRAS, **478**, 3, 2999–3015 (2018a).

Khostovan, A. A., D. Sobral, B. Mobasher, J. Matthee, R. K. Cochrane, N. Chartab Soltani, M. Jafariyazani, A. Paulino-Afonso, S. Santos and J. Calhau, "The clustering of typical Ly$$ emitters from $z 2.5 - 6$: host halo masses depend on Ly$$ and UV luminosities", arXiv e-prints p. arXiv:1811.00556 (2018b).

Kobayashi, M. A. R., T. Totani and M. Nagashima, "Lyα Emitters in Hierarchical Galaxy Formation", ApJ, **670**, 2, 919–927 (2007).

Kobayashi, M. A. R., T. Totani and M. Nagashima, "Lyα Emitters in Hierarchical Galaxy Formation. II. Ultraviolet Continuum Luminosity Function and Equivalent Width Distribution", ApJ, **708**, 2, 1119–1134 (2010).

Konno, A., M. Ouchi, Y. Ono, K. Shimasaku, T. Shibuya, H. Furusawa, K. Nakajima, Y. Naito, R. Momose, S. Yuma and M. Iye, "Accelerated Evolution of the Lyα Luminosity Function at z ¿~7 Revealed by the Subaru Ultra-deep Survey for Lyα Emitters at z = 7.3", ApJ, **797**, 1, 16 (2014).

Konno, A., M. Ouchi, T. Shibuya, Y. Ono, K. Shimasaku, Y. Taniguchi, T. Nagao, M. A. R. Kobayashi, M. Kajisawa, N. Kashikawa, A. K. Inoue, M. Oguri, H. Furusawa, T. Goto, Y. Harikane, R. Higuchi, Y. Komiyama, H. Kusakabe, S. Miyazaki, K. Nakajima and S.-Y. Wang, "SILVERRUSH. IV. Lyα luminosity functions at z = 5.7 and 6.6 studied with ~1300 Lyα emitters on the 14-21 deg^2 sky", PASJ, **70**, S16 (2018).

Kovač, K., R. S. Somerville, J. E. Rhoads, S. Malhotra and J. Wang, "Clustering of Lyα Emitters at z ~ 4.5", ApJ, **668**, 15–22 (2007).

Kreisch, C. D., A. Pisani, F. Villaescusa-Navarro, D. N. Spergel, B. D. Wandelt, N. Hamaus and A. E. Bayer, "The GIGANTES dataset: precision cosmology from voids in the machine learning era", arXiv e-prints p. arXiv:2107.02304 (2021).

Kusakabe, H., K. Shimasaku, M. Ouchi, K. Nakajima, R. Goto, T. Hashimoto, A. Konno, Y. Harikane, J. D. Silverman and P. L. Capak, "The stellar mass, star formation rate and dark matter halo properties of LAEs at z 2", Publications of the Astronomical Society of Japan **70**, 4 (2018).

Landy, S. D. and A. S. Szalay, "Bias and variance of angular correlation functions", ApJ, **412**, 64–71 (1993).

Laursen, P., "IGMtransfer: Intergalactic Radiative Transfer Code", (2011).

Laursen, P., J. Sommer-Larsen and A. O. Razoumov, "Intergalactic Transmission and Its Impact on the Lyα Line", ApJ, **728**, 1, 52 (2011).

Little, B. and D. H. Weinberg, "Cosmic Voids and Biased Galaxy Formation", MNRAS, **267**, 605 (1994).

Madau, P. and F. Haardt, "Cosmic Reionization after Planck: Could Quasars Do It All?", ApJL, **813**, 1, L8 (2015).

Madau, P., F. Haardt and M. J. Rees, "Radiative Transfer in a Clumpy Universe. III. The Nature of Cosmological Ionizing Sources", ApJ, **514**, 2, 648–659 (1999).

Malhotra, S. and J. E. Rhoads, "Large Equivalent Width Lyα line Emission at z=4.5: Young Galaxies in a Young Universe?", ApJ, **565**, L71–L74 (2002).

Malhotra, S. and J. E. Rhoads, "Luminosity Functions of Lyα Emitters at Redshifts z=6.5 and z=5.7: Evidence against Reionization at z=6.5", ApJL, **617**, L5–L8 (2004a).

Malhotra, S. and J. E. Rhoads, "Luminosity Functions of Lyα Emitters at Redshifts z=6.5 and z=5.7: Evidence against Reionization at z¡=6.5", ApJL, **617**, 1, L5–L8 (2004b).

Malhotra, S. and J. E. Rhoads, "The Volume Fraction of Ionized Intergalactic Gas at Redshift z=6.5", ApJL, **647**, 2, L95–L98 (2006).

Malhotra, S., J. E. Rhoads, S. L. Finkelstein, N. Hathi, K. Nilsson, E. McLinden and N. Pirzkal, "Sizing up Lyα and Lyman Break Galaxies", ApJL, **750**, L36 (2012).

Matthee, J., D. Sobral, I. Oteo, P. Best, I. Smail, H. Röttgering and A. Paulino-Afonso, "The CALYMHA survey: Lyα escape fraction and its dependence on galaxy properties at z = 2.23", MNRAS, **458**, 1, 449–467 (2016).

Matthee, J. J. A., D. Sobral, A. M. Swinbank, I. Smail, P. N. Best, J.-W. Kim, M. Franx, B. Milvang-Jensen and J. Fynbo, "A 10 deg^2 Lyman α survey at z=8.8 with spectroscopic follow-up: strong constraints on the luminosity function and implications for other surveys", MNRAS, **440**, 2375–2387 (2014).

Maurogordato, S. and M. Lachieze-Rey, "Void Probabilities in the Galaxy Distribution: Scaling and Luminosity Segregation", ApJ, **320**, 13 (1987).

McGreer, I. D., A. Mesinger and V. D'Odorico, "Model-independent evidence in favour of an end to reionization by z $\approx$ 6", MNRAS, **447**, 1, 499–505 (2015).

McQuinn, M., "The Evolution of the Intergalactic Medium", ARA&A, **54**, 313–362 (2016).

McQuinn, M., L. Hernquist, M. Zaldarriaga and S. Dutta, "Studying reionization with Lyα emitters", MNRAS, **381**, 1, 75–96 (2007).

Mekjian, A. Z., "Generalized Statistical Models of Voids and Hierarchical Structure in Cosmology", ApJ, **655**, 1, 1–10 (2007).

Mesinger, A. and S. R. Furlanetto, "Lyα emitters during the early stages of reionization", MNRAS, **386**, 1990–2002 (2008).

Miralda-Escudé, J., "Reionization of the Intergalactic Medium and the Damping Wing of the Gunn-Peterson Trough", ApJ, **501**, 1, 15–22 (1998).

Morales, A., C. Mason, S. Bruton, M. Gronke, F. Haardt and C. Scarlata, "The Evolution of the Lyman-Alpha Luminosity Function During Reionization", arXiv e-prints p. arXiv:2101.01205 (2021).

Murayama, T., Y. Taniguchi, N. Z. Scoville, M. Ajiki, D. B. Sanders, B. Mobasher, H. Aussel, P. Capak, A. Koekemoer, Y. Shioya, T. Nagao, C. Carilli, R. S. Ellis, B. Garilli, M. Giavalisco, M. G. Kitzbichler, O. Le Fèvre, D. Maccagni, E. Schinnerer, V. Smolčić, S. Tribiano, A. Cimatti, Y. Komiyama, S. Miyazaki, S. S. Sasaki, J. Koda and H. Karoji, "Lyα Emitters at Redshift 5.7 in the COSMOS Field", ApJS, **172**, 1, 523–544 (2007).

Nagamine, K., M. Ouchi, V. Springel and L. Hernquist, "Lyman-α Emitters and Lyman-Break Galaxies at z = 3-6 in Cosmological SPH Simulations", PASJ, **62**, 1455 (2010).

Nagashima, M. and Y. Yoshii, "Hierarchical Formation of Galaxies with Dynamical Response to Supernova-Induced Gas Removal", ApJ, **610**, 1, 23–44 (2004).

Naidu, R. P., S. Tacchella, C. A. Mason, S. Bose, P. A. Oesch and C. Conroy, "Rapid Reionization by the Oligarchs: The Case for Massive, UV-bright, Star-forming Galaxies with High Escape Fractions", ApJ, **892**, 2, 109 (2020).

Nakajima, K., M. Ouchi, K. Shimasaku, Y. Ono, J. C. Lee, S. Foucaud, C. Ly, D. A. Dale, S. Salim, R. Finn, O. Almaini and S. Okamura, "Average Metallicity and Star Formation Rate of Lyα Emitters Probed by a Triple Narrowband Survey", ApJ, **745**, 1, 12 (2012).

Norberg, P., C. M. Baugh, E. Gaztañaga and D. J. Croton, "Statistical analysis of galaxy surveys - I. Robust error estimation for two-point clustering statistics", MNRAS, **396**, 19–38 (2009).

Oesch, P. A., P. G. van Dokkum, G. D. Illingworth, R. J. Bouwens, I. Momcheva, B. Holden, G. W. Roberts-Borsani, R. Smit, M. Franx, I. Labbé, V. González and D. Magee, "A Spectroscopic Redshift Measurement for a Luminous Lyman Break Galaxy at z = 7.730 Using Keck/MOSFIRE", ApJL, **804**, 2, L30 (2015).

Ono, Y., M. Ouchi, B. Mobasher, M. Dickinson, K. Penner, K. Shimasaku, B. J. Weiner, J. S. Kartaltepe, K. Nakajima, H. Nayyeri, D. Stern, N. Kashikawa and H. Spinrad, "Spectroscopic Confirmation of Three z-dropout Galaxies at z = 6.844-7.213: Demographics of Lyα Emission in z ~7 Galaxies", ApJ, **744**, 2, 83 (2012).

Ota, K., M. Iye, N. Kashikawa, K. Shimasaku, M. Kobayashi, T. Totani, M. Nagashima, T. Morokuma, H. Furusawa, T. Hattori, Y. Matsuda, T. Hashimoto and M. Ouchi, "Reionization and Galaxy Evolution Probed by z = 7 Lyα Emitters", ApJ, **677**, 1, 12–26 (2008).

Otto, S., H. D. Politzer, J. Preskill and M. B. Wise, "The Significance of Voids", ApJ, **304**, 62 (1986).

Ouchi, M., Y. Harikane, T. Shibuya, K. Shimasaku, Y. Taniguchi, A. Konno, M. Kobayashi, M. Kajisawa, T. Nagao, Y. Ono, A. K. Inoue, M. Umemura, M. Mori, K. Hasegawa, R. Higuchi, Y. Komiyama, Y. Matsuda, K. Nakajima, T. Saito and S.-Y. Wang, "Systematic Identification of LAEs for Visible Exploration and Reionization Research Using Subaru HSC (SILVERRUSH). I. Program strategy and clustering properties of 2000 Lyα emitters at z = 6-7 over the 0.3-0.5 Gpc² survey area", Publications of the Astronomical Society of Japan **70**, S13 (2018).

Ouchi, M., K. Shimasaku, M. Akiyama, C. Simpson, T. Saito, Y. Ueda, H. Furusawa, K. Sekiguchi, T. Yamada, T. Kodama, N. Kashikawa, S. Okamura, M. Iye, T. Takata, M. Yoshida and M. Yoshida, "The Subaru/XMM-Newton Deep Survey (SXDS). IV. Evolution of Lyα Emitters from z = 3.1 to 5.7 in the 1 deg² Field: Luminosity Functions and AGN", ApJS, **176**, 301–330 (2008).

Ouchi, M., K. Shimasaku, H. Furusawa, M. Miyazaki, M. Doi, M. Hamabe, T. Hayashino, M. Kimura, K. Kodaira, Y. Komiyama, Y. Matsuda, S. Miyazaki, F. Nakata, S. Okamura, M. Sekiguchi, Y. Shioya, H. Tamura, Y. Taniguchi, M. Yagi and N. Yasuda, "Subaru Deep Survey. II. Luminosity Functions and Clustering Properties of Lyα Emitters at z=4.86 in the Subaru Deep Field", ApJ, **582**, 60–68 (2003).

Ouchi, M., K. Shimasaku, H. Furusawa, T. Saito, M. Yoshida, M. Akiyama, Y. Ono, T. Yamada, K. Ota, N. Kashikawa, M. Iye, T. Kodama, S. Okamura, C. Simpson and M. Yoshida, "Statistics of 207 Lyα Emitters at a Redshift Near 7: Constraints on Reionization and Galaxy Formation Models", ApJ, **723**, 869–894 (2010).

Oyarzún, G. A., G. A. Blanc, V. González, M. Mateo and I. Bailey, John I., "A Comprehensive Study of Lyα Emission in the High-redshift Galaxy Population", ApJ, **843**, 2, 133 (2017).

Palunas, P., H. I. Teplitz, P. J. Francis, G. M. Williger and B. E. Woodgate, "The Distribution of Lyα-Emitting Galaxies at z=2.38", ApJ, **602**, 545–554 (2004).

Partridge, R. B. and P. J. E. Peebles, "Are Young Galaxies Visible?", ApJ, **147**, 868 (1967).

Peebles, P. J. E., "Statistical analysis of catalogs of extragalactic objects. VI - The galaxy distribution in the Jagellonian field", ApJ, **196**, 647–651 (1975).

Peebles, P. J. E., *The large-scale structure of the universe* (1980a).

Peebles, P. J. E., *The large-scale structure of the universe* (1980b).

Pentericci, L., A. Fontana, E. Vanzella, M. Castellano, A. Grazian, M. Dijkstra, K. Boutsia, S. Cristiani, M. Dickinson, E. Giallongo, M. Giavalisco, R. Maiolino, A. Moorwood, D. Paris and P. Santini, "Spectroscopic Confirmation of z ~7 Lyman Break Galaxies: Probing the Earliest Galaxies and the Epoch of Reionization", ApJ, **743**, 2, 132 (2011).

Perez, L. A., S. Genel, F. Villaescusa-Navarro, R. S. Somerville, A. Gabrielpillai, D. Anglés-Alcázar, B. D. Wandelt and L. Y. A. Yung, "Constraining cosmology with machine learning and galaxy clustering: the CAMELS-SAM suite", arXiv e-prints p. arXiv:2204.02408 (2022).

Perez, L. A., S. Malhotra, J. E. Rhoads and V. Tilvi, "Void Probability Function of Simulated Surveys of High-redshift Lyα Emitters", ApJ, **906**, 1, 58 (2021).

Pirzkal, N., S. Malhotra, J. E. Rhoads and C. Xu, "Optical-to-Mid-Infrared Observations of Lyα Galaxies at z~5 in the Hubble Ultra Deep Field: A Young and Low-Mass Population", ApJ, **667**, 1, 49–59 (2007).

Pisani, A., E. Massara, D. N. Spergel, D. Alonso, T. Baker, Y.-C. Cai, M. Cautun, C. Davies, V. Demchenko, O. Doré, A. Goulding, M. Habouzit, N. Hamaus, A. Hawken, C. M. Hirata, S. Ho, B. Jain, C. D. Kreisch, F. Marulli, N. Padilla,

G. Pollina, M. Sahlén, R. K. Sheth, R. Somerville, I. Szapudi, R. van de Weygaert, F. Villaescusa-Navarro, B. D. Wandelt and Y. Wang, "Cosmic voids: a novel probe to shed light on our Universe", BAAS, **51**, 3, 40 (2019).

Pisani, A., P. M. Sutter, N. Hamaus, E. Alizadeh, R. Biswas, B. D. Wandelt and C. M. Hirata, "Counting voids to probe dark energy", Phys. Rev. D, **92**, 8, 083531 (2015).

Planck Collaboration, P. A. R. Ade, N. Aghanim, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A. J. Banday, R. B. Barreiro, J. G. Bartlett, N. Bartolo, E. Battaner, R. Battye, K. Benabed, A. Benoît, A. Benoit-Lévy, J. P. Bernard, M. Bersanelli, P. Bielewicz, J. J. Bock, A. Bonaldi, L. Bonavera, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, J. F. Cardoso, A. Catalano, A. Challinor, A. Chamballu, R. R. Chary, H. C. Chiang, J. Chluba, P. R. Christensen, S. Church, D. L. Clements, S. Colombi, L. P. L. Colombo, C. Combet, A. Coulais, B. P. Crill, A. Curto, F. Cuttaia, L. Danese, R. D. Davies, R. J. Davis, P. de Bernardis, A. de Rosa, G. de Zotti, J. Delabrouille, F. X. Désert, E. Di Valentino, C. Dickinson, J. M. Diego, K. Dolag, H. Dole, S. Donzelli, O. Doré, M. Douspis, A. Ducout, J. Dunkley, X. Dupac, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, M. Farhang, J. Fergusson, F. Finelli, O. Forni, M. Frailis, A. A. Fraisse, E. Franceschi, A. Frejsel, S. Galeotta, S. Galli, K. Ganga, C. Gauthier, M. Gerbino, T. Ghosh, M. Giard, Y. Giraud-Héraud, E. Giusarma, E. Gjerløw, J. González-Nuevo, K. M. Górski, S. Gratton, A. Gregorio, A. Gruppuso, J. E. Gudmundsson, J. Hamann, F. K. Hansen, D. Hanson, D. L. Harrison, G. Helou, S. Henrot-Versillé, C. Hernández-Monteagudo, D. Herranz, S. R. Hildebrandt, E. Hivon, M. Hobson, W. A. Holmes, A. Hornstrup, W. Hovest, Z. Huang, K. M. Huffenberger, G. Hurier, A. H. Jaffe, T. R. Jaffe, W. C. Jones, M. Juvela, E. Keihänen, R. Keskitalo, T. S. Kisner, R. Kneissl, J. Knoche, L. Knox, M. Kunz, H. Kurki-Suonio, G. Lagache, A. Lähteenmäki, J. M. Lamarre, A. Lasenby, M. Lattanzi, C. R. Lawrence, J. P. Leahy, R. Leonardi, J. Lesgourgues, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, M. Linden-Vørnle, M. López-Caniego, P. M. Lubin, J. F. Macías-Pérez, G. Maggio, D. Maino, N. Mandolesi, A. Mangilli, A. Marchini, M. Maris, P. G. Martin, M. Martinelli, E. Martínez-González, S. Masi, S. Matarrese, P. McGehee, P. R. Meinhold, A. Melchiorri, J. B. Melin, L. Mendes, A. Mennella, M. Migliaccio, M. Millea, S. Mitra, M. A. Miville-Deschênes, A. Moneti, L. Montier, G. Morgante, D. Mortlock, A. Moss, D. Munshi, J. A. Murphy, P. Naselsky, F. Nati, P. Natoli, C. B. Netterfield, H. U. Nørgaard-Nielsen, F. Noviello, D. Novikov, I. Novikov, C. A. Oxborrow, F. Paci, L. Pagano, F. Pajot, R. Paladini, D. Paoletti, B. Partridge, F. Pasian, G. Patanchon, T. J. Pearson, O. Perdereau, L. Perotto, F. Perrotta, V. Pettorino, F. Piacentini, M. Piat, E. Pierpaoli, D. Pietrobon, S. Plaszczynski, E. Pointecouteau, G. Polenta, L. Popa, G. W. Pratt, G. Prézeau, S. Prunet, J. L. Puget, J. P. Rachen, W. T. Reach, R. Rebolo, M. Reinecke, M. Remazeilles, C. Renault, A. Renzi, I. Ristorcelli, G. Rocha, C. Rosset, M. Rossetti, G. Roudier, B. Rouillé d'Orfeuil, M. Rowan-Robinson, J. A. Rubiño-Martín, B. Rusholme, N. Said, V. Salvatelli, L. Salvati, M. Sandri, D. Santos, M. Savelainen, G. Savini, D. Scott, M. D. Seiffert, P. Serra, E. P. S. Shellard, L. D. Spencer, M. Spinelli, V. Stolyarov, R. Stompor, R. Sudiwala, R. Sunyaev, D. Sutton, A. S. Suur-Uski, J. F. Sygnet, J. A. Tauber, L. Terenzi, L. Toffo-

latti, M. Tomasi, M. Tristram, T. Trombetti, M. Tucci, J. Tuovinen, M. Türler, G. Umana, L. Valenziano, J. Valiviita, F. Van Tent, P. Vielva, F. Villa, L. A. Wade, B. D. Wandelt, I. K. Wehus, M. White, S. D. M. White, A. Wilkinson, D. Yvon, A. Zacchei and A. Zonca, "Planck 2015 results. XIII. Cosmological parameters", A&A, **594**, A13 (2016).

Reddy, N. A. and C. C. Steidel, "A Steep Faint-End Slope of the UV Luminosity Function at z ~2-3: Implications for the Global Stellar Mass Density and Star Formation in Low-Mass Halos", ApJ, **692**, 1, 778–803 (2009).

Repp, A. and I. Szapudi, "Galaxy Bias and σ_8 from Counts in Cells from the SDSS Main Sample", arXiv e-prints p. arXiv:2006.01146 (2020).

Rhoads, J. E., "Testing the Topology of Reionization", arXiv e-prints p. arXiv:0708.2909 (2007).

Rhoads, J. E. and S. Malhotra, "Lyα Emitters at Redshift z = 5.7", ApJL, **563**, 1, L5–L9 (2001).

Rhoads, J. E., S. Malhotra, A. Dey, D. Stern, H. Spinrad and B. T. Jannuzi, "First Results from the Large-Area Lyman Alpha Survey", ApJL, **545**, L85–L88 (2000).

Rhoads, J. E., S. Malhotra, D. Stern, M. Dickinson, N. Pirzkal, H. Spinrad, N. Reddy, N. Hathi, N. Grogin, A. Koekemoer, M. A. Peth, S. Cohen, Z. Zheng, T. Budavari, I. Ferreras, J. P. Gardner, C. Gronwall, Z. Haiman, M. Kümmel, G. Meurer, L. Moustakas, N. Panagia, A. Pasquali, K. Sahu, S. di Serego Alighieri, R. Somerville, A. Straughn, J. Walsh, R. Windhorst, C. Xu and H. Yan, "A Lyman Break Galaxy in the Epoch of Reionization from Hubble Space Telescope Grism Spectroscopy", ApJ, **773**, 1, 32 (2013).

Robertson, B. E., R. S. Ellis, J. S. Dunlop, R. J. McLure and D. P. Stark, "Early star-forming galaxies and the reionization of the Universe", Nature, **468**, 7320, 49–55 (2010).

Robertson, B. E., S. R. Furlanetto, E. Schneider, S. Charlot, R. S. Ellis, D. P. Stark, R. J. McLure, J. S. Dunlop, A. Koekemoer, M. A. Schenker, M. Ouchi, Y. Ono, E. Curtis-Lake, A. B. Rogers, R. A. A. Bowler and M. Cirasuolo, "New Constraints on Cosmic Reionization from the 2012 Hubble Ultra Deep Field Campaign", ApJ, **768**, 1, 71 (2013).

Ryden, B. S. and A. L. Melott, "Voids in Real Space and in Redshift Space", ApJ, **470**, 160 (1996).

Santos, M. R., "Probing reionization with Lyman α emission lines", MNRAS, **349**, 3, 1137–1152 (2004).

Santos, S., D. Sobral and J. Matthee, "The Lyα luminosity function at z = 5.7 - 6.6 and the steep drop of the faint end: implications for reionization", MNRAS, **463**, 1678–1691 (2016).

Saslaw, W. C. and F. Fang, "The Thermodynamic Description of the Cosmological Many-Body Problem", ApJ, **460**, 16 (1996).

Saslaw, W. C. and A. J. S. Hamilton, "Thermodynamics and galaxy clustering - Nonlinear theory of high order correlations", ApJ, **276**, 13–25 (1984).

Schenker, M. A., D. P. Stark, R. S. Ellis, B. E. Robertson, J. S. Dunlop, R. J. McLure, J.-P. Kneib and J. Richard, "Keck Spectroscopy of Faint $3 < z < 8$ Lyman Break Galaxies: Evidence for a Declining Fraction of Emission Line Sources in the Redshift Range $6 < z < 8$", ApJ, **744**, 2, 179 (2012).

Sheth, R. K., "Press-Schechter, thermodynamics and gravitational clustering", MNRAS, **274**, 1, 213–220 (1995).

Sheth, R. K., "An excursion set model for the distribution of dark matter and dark matter haloes", MNRAS, **300**, 4, 1057–1070 (1998).

Shimasaku, K., T. Hayashino, Y. Matsuda, M. Ouchi, K. Ohta, S. Okamura, H. Tamura, T. Yamada and R. Yamauchi, "Large Cosmic Variance in the Clustering Properties of Lyα Emitters at z ˜$= 5$", ApJ, **605**, L93–L96 (2004).

Shimasaku, K., N. Kashikawa, M. Doi, C. Ly, M. A. Malkan, Y. Matsuda, M. Ouchi, T. Hayashino, M. Iye, K. Motohara, T. Murayama, T. Nagao, K. Ohta, S. Okamura, T. Sasaki, Y. Shioya and Y. Taniguchi, "Lyα Emitters at z $= 5.7$ in the Subaru Deep Field", Publications of the Astronomical Society of Japan **58**, 313–334 (2006).

Shioya, Y., Y. Taniguchi, S. S. Sasaki, T. Nagao, T. Murayama, T. Saito, Y. Ideue, A. Nakajima, K. Matsuoka, J. Trump, N. Z. Scoville, D. B. Sand ers, B. Mobasher, H. Aussel, P. Capak, J. Kartaltepe, A. Koekemoer, C. Carilli, R. S. Ellis, B. Garilli, M. Giavalisco, M. G. Kitzbichler, C. Impey, O. LeFevre, E. Schinnerer and V. Smolcic, "Photometric Properties of Lyα Emitters at z ≈ 4.86 in the COSMOS 2 Square Degree Field", ApJ, **696**, 1, 546–561 (2009).

Sinha, M. and L. H. Garrison, "CORRFUNC - a suite of blazing fast correlation functions on the CPU", MNRAS, **491**, 2, 3022–3041 (2020).

Skibba, R. A., M. S. M. Smith, A. L. Coil, J. Moustakas, J. Aird, M. R. Blanton, A. D. Bray, R. J. Cool, D. J. Eisenstein, A. J. Mendez, K. C. Wong and G. Zhu, "PRIMUS: Galaxy Clustering as a Function of Luminosity and Color at 0.2 ¡ z ¡ 1", ApJ, **784**, 2, 128 (2014).

Sobacchi, E. and A. Mesinger, "The clustering of Lyman α emitters at z ≈ 7: implications for reionization and host halo masses", MNRAS, **453**, 2, 1843–1854 (2015a).

Sobacchi, E. and A. Mesinger, "The clustering of Lyman α emitters at z ≈ 7: implications for reionization and host halo masses", MNRAS, **453**, 2, 1843–1854 (2015b).

Sobral, D., J. Matthee, P. Best, A. Stroe, H. Röttgering, I. Oteo, I. Smail, L. Morabito and A. Paulino-Afonso, "The CALYMHA survey: Lyα luminosity function and global escape fraction of Lyα photons at z $= 2.23$", MNRAS, **466**, 1, 1242–1258 (2017).

Sobral, D., J. Matthee, B. Darvish, D. Schaerer, B. Mobasher, H. J. A. Röttgering, S. Santos and S. Hemmati, "Evidence for PopIII-like Stellar Populations in the Most Luminous Lyman-α Emitters at the Epoch of Reionization: Spectroscopic Confirmation", ApJ, **808**, 2, 139 (2015).

Sobral, D., S. Santos, J. Matthee, A. Paulino-Afonso, B. Ribeiro, J. Calhau and A. A. Khostovan, "Slicing COSMOS with SC4K: the evolution of typical Ly α emitters and the Ly α escape fraction from z 2 to 6", MNRAS, **476**, 4725–4752 (2018).

Somerville, R. S., P. F. Hopkins, T. J. Cox, B. E. Robertson and L. Hernquist, "A semi-analytic model for the co-evolution of galaxies, black holes and active galactic nuclei", MNRAS, **391**, 2, 481–506 (2008).

Somerville, R. S., G. Popping and S. C. Trager, "Star formation in semi-analytic galaxy formation models with multiphase gas", MNRAS, **453**, 4, 4337–4367 (2015).

Somerville, R. S. and J. R. Primack, "Semi-analytic modelling of galaxy formation: the local Universe", MNRAS, **310**, 4, 1087–1110 (1999).

Springel, V., "The cosmological simulation code GADGET-2", MNRAS, **364**, 1105–1134 (2005).

Stark, D. P., R. S. Ellis, K. Chiu, M. Ouchi and A. Bunker, "Keck spectroscopy of faint 3 ¡ z ¡ 7 Lyman break galaxies - I. New constraints on cosmic reionization from the luminosity and redshift-dependent fraction of Lyman α emission", MNRAS, **408**, 3, 1628–1648 (2010).

Sutter, P. M., G. Lavaux, N. Hamaus, A. Pisani, B. D. Wandelt, M. Warren, F. Villaescusa-Navarro, P. Zivick, Q. Mao and B. B. Thompson, "VIDE: The Void IDentification and Examination toolkit", Astronomy and Computing **9**, 1–9 (2015).

Szapudi, I., "A New Method for Calculating Counts in Cells", ApJ, **497**, 16–20 (1998).

Taniguchi, Y., M. Ajiki, T. Nagao, Y. Shioya, T. Murayama, N. Kashikawa, K. Kodaira, N. Kaifu, H. Ando, H. Karoji, M. Akiyama, K. Aoki, M. Doi, S. S. Fujita, H. Furusawa, T. Hayashino, F. Iwamuro, M. Iye, N. Kobayashi, T. Kodama, Y. Komiyama, Y. Matsuda, S. Miyazaki, Y. Mizumoto, T. Morokuma, K. Motohara, K. Nariai, K. Ohta, Y. Ohyama, S. Okamura, M. Ouchi, T. Sasaki, Y. Sato, K. Sekiguchi, K. Shimasaku, H. Tamura, M. Umemura, T. Yamada, N. Yasuda and M. Yoshida, "The SUBARU Deep Field Project: Lymanα Emitters at a Redshift of 6.6", Publications of the Astronomical Society of Japan **57**, 165–182 (2005).

Thacker, R. J. and H. M. P. Couchman, "A parallel adaptive P^{3}M code with hierarchical particle reordering", Computer Physics Communications **174**, 540–554 (2006).

Tilvi, V., S. Malhotra, J. E. Rhoads, A. Coughlin, Z. Zheng, S. L. Finkelstein, S. Veilleux, B. Mobasher, J. Wang, R. Probst, R. Swaters, P. Hibon, B. Joshi, J. Zabl, T. Jiang, J. Pharo and H. Yang, "Onset of Cosmic Reionization: Evidence of an Ionized Bubble Merely 680 Myr after the Big Bang", ApJL, **891**, 1, L10 (2020).

Tilvi, V., S. Malhotra, J. E. Rhoads, E. Scannapieco, R. J. Thacker, I. T. Iliev and G. Mellema, "A Physical Model of Lyα Emitters", ApJ, **704**, 724–732 (2009).

Tilvi, V., C. Papovich, S. L. Finkelstein, J. Long, M. Song, M. Dickinson, H. C. Ferguson, A. M. Koekemoer, M. Giavalisco and B. Mobasher, "Rapid Decline of Lyα Emission toward the Reionization Era", ApJ, **794**, 5 (2014).

Tilvi, V., J. E. Rhoads, P. Hibon, S. Malhotra, J. Wang, S. Veilleux, R. Swaters, R. Probst, H. Krug, S. L. Finkelstein and M. Dickinson, "The Luminosity Function of Lyα Emitters at Redshift z = 7.7", ApJ, **721**, 2, 1853–1860 (2010).

Tinker, J. L., C. Conroy, P. Norberg, S. G. Patiri, D. H. Weinberg and M. S. Warren, "Void Statistics in Large Galaxy Redshift Surveys: Does Halo Occupation of Field Galaxies Depend on Environment?", ApJ, **686**, 53–71 (2008).

Tinker, J. L., D. H. Weinberg and M. S. Warren, "Cosmic Voids and Galaxy Bias in the Halo Occupation Framework", ApJ, **647**, 737–752 (2006).

Totsuji, H. and T. Kihara, "The Correlation Function for the Distribution of Galaxies", Publications of the Astronomical Society of Japan **21**, 221 (1969).

Trainor, R. F., A. L. Strom, C. C. Steidel, G. C. Rudie, Y. Chen and R. L. Theios, "Predicting Lyα Emission from Galaxies via Empirical Markers of Production and Escape in the KBSS", ApJ, **887**, 1, 85 (2019).

Trenti, M. and M. Stiavelli, "Cosmic Variance and Its Effect on the Luminosity Function Determination in Deep High-z Surveys", ApJ, **676**, 2, 767–780 (2008).

Uhlemann, C., O. Friedrich, F. Villaescusa-Navarro, A. Banerjee and S. r. Codis, "Fisher for complements: extracting cosmology and neutrino mass from the counts-in-cells PDF", MNRAS, **495**, 4, 4006–4027 (2020).

Villaescusa-Navarro, F., C. Hahn, E. Massara, A. Banerjee, A. M. Delgado, D. K. Ramanah, T. Charnock, E. Giusarma, Y. Li, E. Allys, A. Brochard, C. Uhlemann, C.-T. Chiang, S. He, A. Pisani, A. Obuljen, Y. Feng, E. Castorina, G. Contardo, C. D. Kreisch, A. Nicola, J. Alsing, R. Scoccimarro, L. Verde, M. Viel, S. Ho, S. Mallat, B. Wandelt and D. N. Spergel, "The Quijote Simulations", ApJS, **250**, 1, 2 (2020).

Walsh, K. and J. Tinker, "Probing Galaxy assembly bias in BOSS galaxies using void probabilities", MNRAS, **488**, 1, 470–479 (2019).

Wang, K., Y.-Y. Mao, A. R. Zentner, F. C. van den Bosch, J. U. Lange, C. M. Schafer, A. S. Villarreal, A. P. Hearin and D. Campbell, "How to optimally constrain galaxy assembly bias: supplement projected correlation functions with count-in-cells statistics", MNRAS, **488**, 3, 3541–3567 (2019).

Wen, D., A. J. Kemball and W. C. Saslaw, "Halo Counts-in-cells for Cosmological Models with Different Dark Energy", arXiv e-prints p. arXiv:2001.06119 (2020).

White, S. D. M., "The hierarchy of correlation functions and its relation to other measures of galaxy clustering", MNRAS, **186**, 145–154 (1979).

Wold, I. G. B., S. Malhotra, J. Rhoads, J. Wang, W. Hu, L. A. Perez, Z.-Y. Zheng, A. A. Khostovan, A. R. Walker, L. F. Barrientos, J. González-López, S. Harish, L. Infante, C. Jiang, J. Pharo, C. Moya-Sierralta, F. E. Bauer, G. Galaz, F. Valdes and H. Yang, "LAGER Lyα Luminosity Function at z 7: Implications for Reionization", ApJ, **927**, 1, 36 (2022).

Wold, I. G. B., S. Malhotra, J. Rhoads, J. Wang, W. Hu, L. A. Perez, Z.-Y. Zheng, A. A. Khostovan, A. R. Walker, L. F. Barrientos, J. González-López, S. Harish, L. Infante, C. Jiang, J. Pharo, C. Moya-Sierralta, F. Valdes and H. Yang, "LAGER Lyα Luminosity Function at $z \sim 7$, Implications for Reionization", arXiv e-prints p. arXiv:2105.12191 (2021).

Wolk, M., H. J. McCracken, S. Colombi, J. N. Fry, M. Kilbinger, P. Hudelot, Y. Mellier and O. Ilbert, "Evolution of hierarchical clustering in the CFHTLS-Wide since z 1", MNRAS, **435**, 1, 2–17 (2013).

Yang, A. and W. C. Saslaw, "The Galaxy Counts-in-cells Distribution from the Sloan Digital Sky Survey", ApJ, **729**, 2, 123 (2011).

Yang, H., L. Infante, J. E. Rhoads, W. Hu, Z. Zheng, S. Malhotra, J. Wang, L. F. Barrientos, W. Kang and C. Jiang, "Lyα Galaxies in the Epoch of Reionization (LAGER): Spectroscopic Confirmation of Two Redshift $\sim$7.0 Galaxies", ApJ, **876**, 2, 123 (2019).

Yung, L. Y. A., R. S. Somerville, S. L. Finkelstein, G. Popping, R. Davé, A. Venkatesan, P. Behroozi and H. C. Ferguson, "Semi-analytic forecasts for JWST - IV. Implications for cosmic reionization and LyC escape fraction", MNRAS, **496**, 4, 4574–4592 (2020).

Zehavi, I., M. R. Blanton, J. A. Frieman, D. H. Weinberg, H. J. Mo, M. A. Strauss, S. F. Anderson, J. Annis, N. A. Bahcall, M. Bernardi, J. W. Briggs, J. Brinkmann, S. Burles, L. Carey, F. J. Castander, A. J. Connolly, I. Csabai, J. J. Dalcanton, S. Dodelson, M. Doi, D. Eisenstein, M. L. Evans, D. P. Finkbeiner, S. Friedman, M. Fukugita, J. E. Gunn, G. S. Hennessy, R. B. Hindsley, Ž. Ivezić, S. Kent, G. R. Knapp, R. Kron, P. Kunszt, D. Q. Lamb, R. F. Leger, D. C. Long, J. Loveday, R. H. Lupton, T. McKay, A. Meiksin, A. Merrelli, J. A. Munn, V. Narayanan, M. Newcomb, R. C. Nichol, R. Owen, J. Peoples, A. Pope, C. M. Rockosi, D. Schlegel, D. P. Schneider, R. Scoccimarro, R. K. Sheth, W. Siegmund, S. Smee, Y. Snir, A. Stebbins, C. Stoughton, M. SubbaRao, A. S. Szalay, I. Szapudi, M. Tegmark, D. L. Tucker, A. Uomoto, D. Vanden Berk, M. S. Vogeley, P. Waddell, B. Yanny and D. G. York, "Galaxy Clustering in Early Sloan Digital Sky Survey Redshift Data", ApJ, **571**, 1, 172–190 (2002).

Zhang, G., Z. Li, J. Liu, D. N. Spergel, C. D. Kreisch, A. Pisani and B. D. Wandelt, "Void halo mass function: A promising probe of neutrino mass", Phys. Rev. D, **102**, 8, 083537 (2020).

Zheng, Z., R. Cen, H. Trac and J. Miralda-Escudé, "Radiative Transfer Modeling of Lyα Emitters. I. Statistics of Spectra and Luminosity", ApJ, **716**, 1, 574–598 (2010).

Zheng, Z.-Y., S. Malhotra, J. E. Rhoads, S. L. Finkelstein, J.-X. Wang, C.-Y. Jiang and Z. Cai, "Lyα Emitter Galaxies at z 2.8 in the Extended Chandra Deep Field South. I. Tracing the Large-scale Structure via Lyα Imaging", The Astrophysical Journal Supplement Series **226**, 23 (2016).

Zheng, Z.-Y., J. E. Rhoads, J.-X. Wang, S. Malhotra, A. Walker, T. Mooney, C. Jiang, W. Hu, P. Hibon, L. Jiang, L. Infante, L. F. Barrientos, G. Galaz, F. Valdes, W. Wester, H. Yang, A. Coughlin, S. Harish, W. Kang, A. A. Khostovan, X. Kong, L. A. Perez, J. Pharo, I. Wold and X. Zheng, "Design for the First Narrowband Filter for the Dark Energy Camera: Optimizing the LAGER Survey for z 7 Galaxies", PASP, **131**, 1001, 074502 (2019).

Zheng, Z.-Y., J. Wang, J. Rhoads, L. Infante, S. Malhotra, W. Hu, A. R. Walker, L. Jiang, C. Jiang, P. Hibon, A. Gonzalez, X. Kong, X. Zheng, G. Galaz and L. F. Barrientos, "First Results from the Lyman Alpha Galaxies in the Epoch of Reionization (LAGER) Survey: Cosmological Reionization at z 7", ApJL, **842**, 2, L22 (2017).

Zheng, Z.-Y., J.-X. Wang, S. Malhotra, J. E. Rhoads, S. L. Finkelstein and K. Finkelstein, "Lyα equivalent width distribution of Lyα emitting galaxies at redshift z ∼ 4.5", MNRAS, **439**, 1, 1101–1109 (2014).

Zitrin, A., I. Labbé, S. Belli, R. Bouwens, R. S. Ellis, G. Roberts-Borsani, D. P. Stark, P. A. Oesch and R. Smit, "Lymanα Emission from a Luminous z = 8.68 Galaxy: Implications for Galaxies as Tracers of Cosmic Reionization", ApJL, **810**, 1, L12 (2015).

APPENDIX A

PAPER I: HIERARCHICAL SCALING MODELS

A   PAPER I: HIERARCHICAL SCALING MODELS

As published with Paper I, Perez *et al.* (2021).

We summarize the various models of hierarchical scaling between the volume-averaged correlation functions that we test for our simulated LAE catalogs. We discuss the negative binomial model in great detail in §2.4.3. The bulk of this list and the functional forms of the models come from C05, Cr04, and Fry and Colombi (2013). The order of these models follows the order of their curves in Figure 2.4, from top to bottom.

Thermodynamic & Gravitational Quasi-Equilibrium Models

(The naming schemes for these models have evolved over time, and what we call the QEM is referred to as the thermodynamic model in several places. We identify the QEM as Fry and Colombi 2013 do, and call the final form of the model that Fry 1986 derives the thermodynamic model.)

The thermodynamic (black dash-dot curve) and the so-called gravitational quasi-equilibrium (black crosses) in 2.4 are both derived from the same occupation probability distribution of Saslaw and Hamilton (1984). They developed a thermodynamic theory of the properties of gravitational clustering that yields this occupation distribution (Fry, 1986):

$$P_N = \frac{\bar{N}(1-b)}{N!}[\bar{n}(1-b) + Nb]^{N-1}\exp(-\bar{N}(1-b) - Nb). \tag{1}$$

A thorough explanation of the derivation of the two models can be found in Fry (1986), which we summarize here. The "quasi-equilibrium" model (black crosses in Figure 2.4) comes from approximation discrete realization of a continuous background number density:

$$\chi_{\mathrm{QEM}} = (1 + \bar{N}\bar{\xi})^{-1/2} \tag{2}$$

One can then apply a continuum limit to process the P_N distribution into a discretized continuum for large $\bar{N}$:

$$N^{N-2}Q_N = (2N-3)!!, \quad K = \bar{\xi}^{-1}[1 - (1 - 2\bar{N}\bar{\xi})^{1/2}], \tag{3}$$

which gives the form of the "thermodynamic" model we plot as a black dash-dot curve in 2.4:

$$\chi_{\mathrm{thermo}} = \left((1 + 2\bar{N}\bar{\xi})^{1/2} - 1\right)/\bar{N}\bar{\xi} \tag{4}$$

Ahmad *et al.* (2002) derived the thermodynamic model from statistical mechanics, imagining that galaxy clustering evolves through a sequence of quasi-equilibrium states and supplying fundamental justification for the model. Sheth (1998) derived a model that yields the same S_p scaling coefficients as the thermodynamic model, but by instead approaching the evolution of the dark matter halo mass function as an extension of the excusion set model. Fry (1985) details that the thermodynamic

136

model is only valid in large volumes (once $\bar{N}\bar{\xi}$ has converged), and expresses concern at how such large scales could have become thermodynamically relaxed over age of universe. (Given that at the redshifts of our simulated LAEs, the universe was 800 million to 2.1 billion years old, this concern remains.)

Geometric Model

This model begins with an occupancy probability distribution $p_n \propto p^n$, and is a specialized case of the hyper-geometric model of Mekjian (2007), the Hamilton (1988) model, and alternate form of the model in Alimi *et al.* (1990). The geometric hierarchical model takes the form:

$$\chi_{\text{geom}} = 1/(1 + \frac{1}{2}\bar{N}\bar{\xi})$$

(5)

Minimal Model

The minimal model can be derived by taking the moments of galaxy counts in a hierarchical Poisson model, in which galaxies form in randomly places clusters of N_c galaxies and $\bar{N}\bar{\xi} \gg 1$ (Fry, 1988). This model is another limiting case of the Mekjian (2007) model. The minimal Poisson model has all scaling coefficients $S_p = 1$ for all p, and takes the form:

$$\chi_{\text{min}} = \left(1 - \exp(-\bar{N}\bar{\xi})\right)/\bar{N}\bar{\xi}$$

(6)

APPENDIX B

PAPER III: CONSTRAINING REIONIZATION WITH ROMAN

B PAPER III: CONSTRAINING REIONIZATION WITH ROMAN

Detailed VPF Distributions

What do the distributions of the VPF across all the independent $\Delta z = 0.2$ slices look like for each redshift-reionization scenario? We compare the distribution of the VPF measurements across: the 'full-face' slices of approximately 13-16 deg^2 (top row), the few dozen slices of approximately 4 deg^2 (middle row), and the several dozen slices of approximately 1 deg^2 (bottom row) in Figures 1 - 3. Columns indicate the J14 volume analyzed in the scenario, with the most neutral to the left and increasing in ionization fraction. Each figure shows, in a solid line whose color corresponds to the $\langle x_i \rangle_v$ value of the simulation (e.g. in Figure 4.1, the mean of the VPF across the several full-face slices). The histograms for the 4 deg^2 and 1 deg^2 slices also show the 1σ standard deviation across the full face slices. Increasing the survey area steeply narrows the distribution of sampled slices about what is likely the true VPF value.

The histograms are not normalized, and we use (5, 15, 20) default-MATPLOTLIB bins for the (~ 16, ~ 4, ~ 1) deg^2 slices. We also remind readers that each slices' VPF measurement is the average across 5 calculations of the VPF with the Banerjee and Abel (2020) k-NN method. Though not shown here, all valid distributions are clustered compared to random (e.g. Paper 1 Figures). Finally, a quirk of the VPF is that the spread of a distribution is related to its central value: more clustered VPFs (closer to zero) also have narrower distributions.
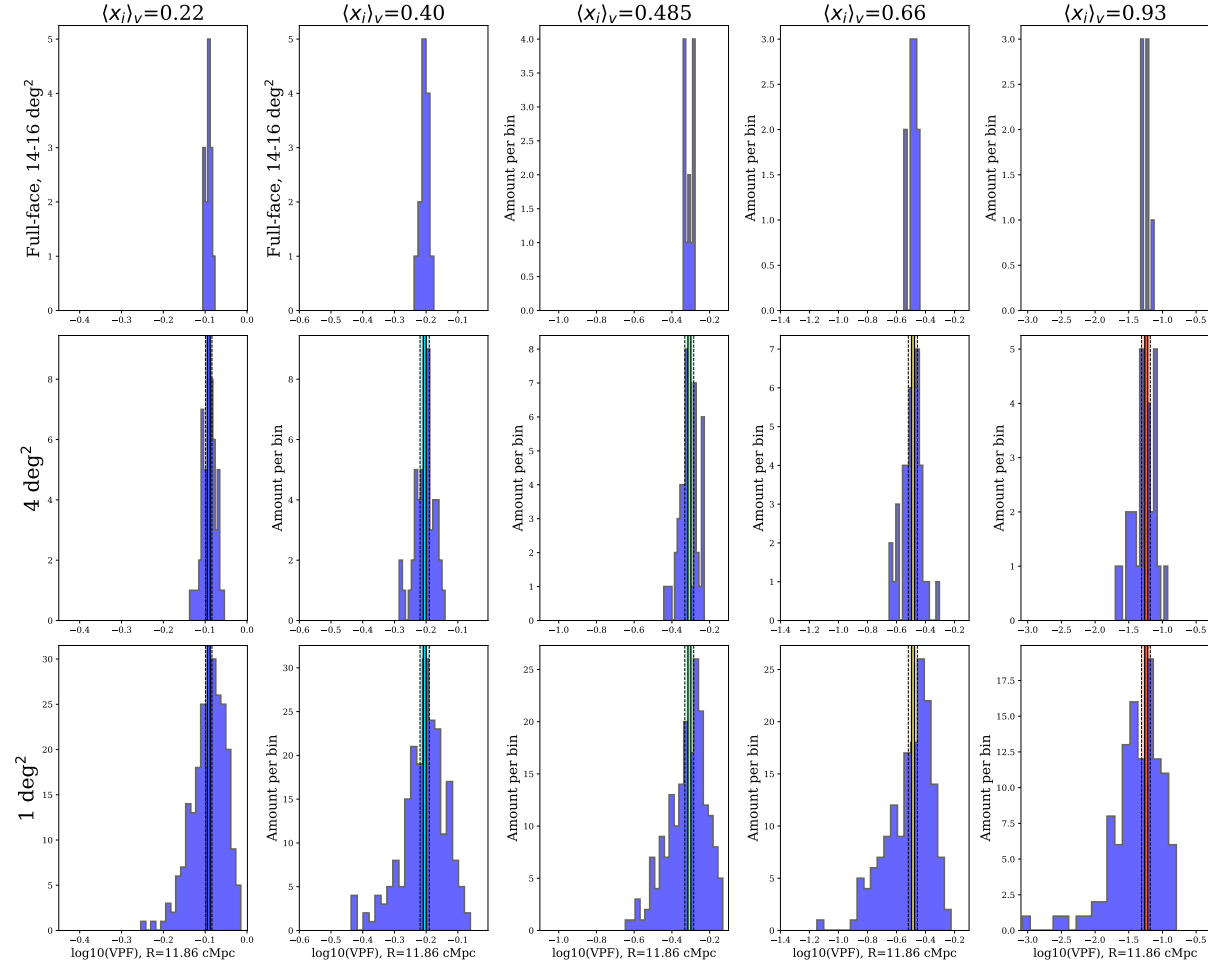
Figure 1: Focused distributions of the VPF measured at $R = 11.86$ cMpc for the Finkelstein *et al.* (2019) redshift-reionization scenarios; the columns are the J14 simulation volume (left: $\langle x_i \rangle_v$=0.22, in increasing order until right: $\langle x_i \rangle$=0.93). The rows are for the Roman survey area probed: full-face 13-16 deg^2 (top), ~ 4 deg^2 (middle), ~ 1 deg^2 (bottom). The solid lines in all figures indicate the mean VPF across the full-face slices; the semi-transparent shading in the ~ 4 deg^2 and ~ 1 deg^2 histograms are the 1σ standard deviation measured on the full-face distributions.
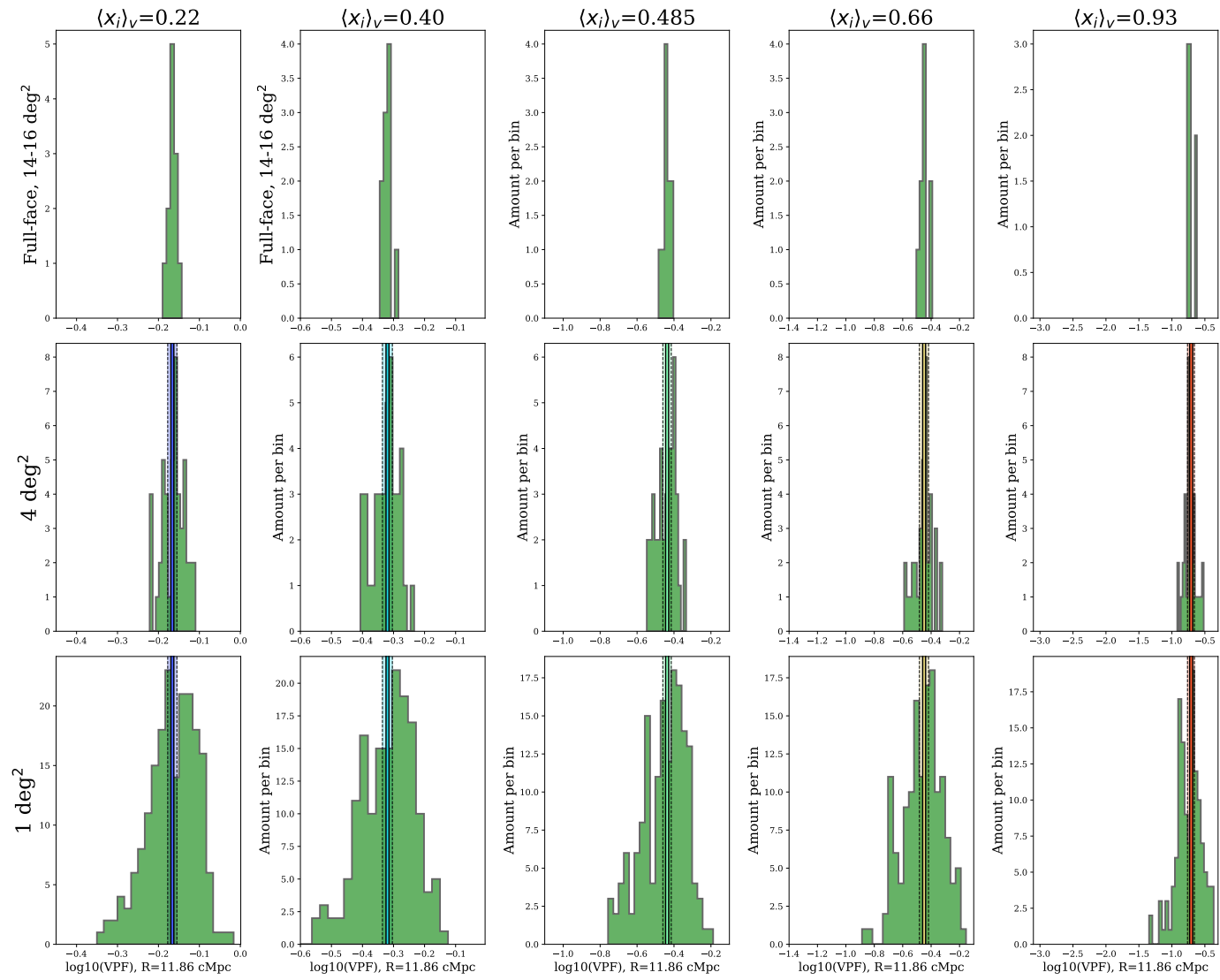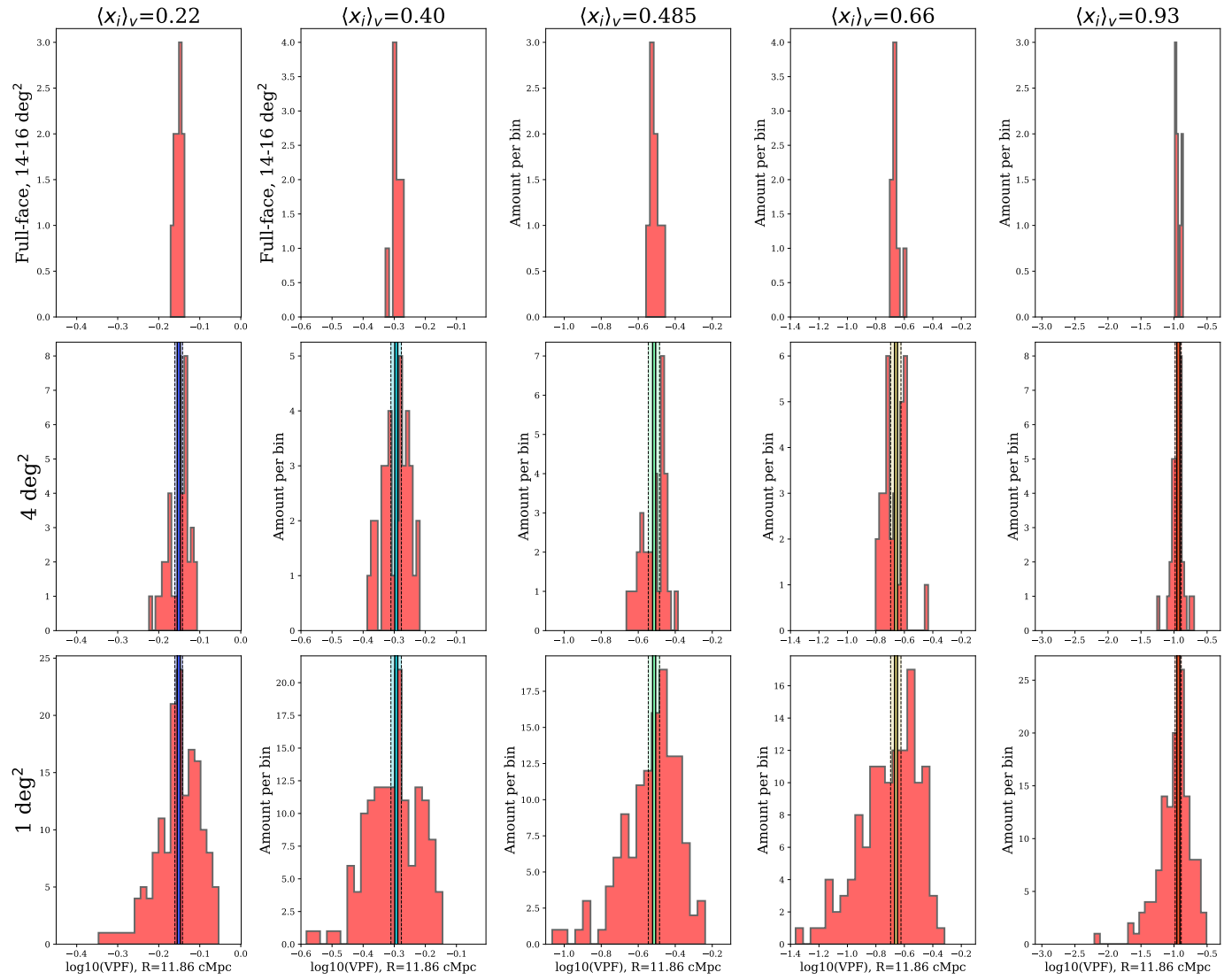
Figure 2: Like Figure 1, but for Yung *et al.* (2020).

Figure 3: Like Figure 1, but for Naidu *et al.* (2020).

APPENDIX C

CO-AUTHOR APPROVALS FOR PUBLISHED WORKS

ASU

Lucia Perez <laperez5@asu.edu>

IMPORTANT: Permission to put our 2021 LAE VPF paper in my dissertation?
3 messages

Lucia Perez <laperez5@asu.edu>   Fri, Apr 22, 2022 at 1:28 PM
To: Sangeeta Malhotra <sangeeta.malhotra@asu.edu>, James Rhoads <James.Rhoads@asu.edu>

Hi Sangeeta and James,
As part of the PhD process, ASU requires written permission to use published papers in dissertations (attached is the policy).
Can I have your permission to put our paper "Void Probability Function of Simulated Surveys of High-redshift Lyα Emitters" (Perez, Malhotra, Rhoads, & Tilvi; The Astrophysical Journal, Volume 906, Issue 1, id.58; January 2021) in my dissertation? A response to this email is plenty.
Thank you so much! We're doing it!!!!
Best,
Lucia

--
Lucia A. Perez
School for Earth and Space Exploration
PhD *Recipient*, Astrophysics

📄 **policy-on-using-previously-published-work.pdf**
133K

Sangeeta Malhotra <Sangeeta.Malhotra@asu.edu>   Fri, Apr 22, 2022 at 2:05 PM
To: Lucia Perez <laperez5@asu.edu>

You have my permission to use this paper as a chapter of your thesis.

Sangeeta

[Quoted text hidden]

James Rhoads <James.Rhoads@asu.edu>   Fri, Apr 22, 2022 at 3:15 PM
To: Lucia Perez <laperez5@asu.edu>
Cc: Sangeeta Malhotra <Sangeeta.Malhotra@asu.edu>

Hi Lucia.

You have my permission to use the paper you mentioned as part of youe PhD thesis.

Regards, James Rhoads
[Quoted text hidden]

Figure 4: Permission for Paper I, from S. Malhotra and J.E. Rhoads.

C   CO-AUTHOR APPROVALS FOR PUBLISHED WORKS

Written, express permission from all co-authors for published, submitted, and to-be-submitted works presented in Lucia A. Perez's dissertation.

ASU

Lucia Perez <laperez5@asu.edu>

IMPORTANT: Permission to put our LAE VPF paper in my dissertation?

V. Tilvi <tilvi@asu.edu>   Wed, Mar 23, 2022 at 12:53 PM
To: Lucia Perez <laperez5@asu.edu>

Hi Lucia,

Great news !

Yes, for sure you have my permission for this.

Cheers,
Tilvi

> On Mar 23, 2022, at 8:25 PM, Lucia Perez <laperez5@asu.edu> wrote:
>
> Hi Tilvi!
> I hope you've been well! I'm FINALLY getting everything ready to defend and graduate! Currently scrambling to finish up some work,
> compile my dissertation, and schedule the defense. As you may be familiar with from working with a bunch of us throughout the last few
> years, ASU wants written permission to use published papers in dissertations (attached is the policy).
> Can I have your permission to put our paper"Void Probability Function of Simulated Surveys of High-redshift Lyα Emitters" (Perez, Malhotra,
> Rhoads, & Tilvi; The Astrophysical Journal, Volume 906, Issue 1, id.58; January 2021) in my dissertation? A response to this email is plenty.
> Thank you so much! I'll also share an invitation to my defense once it's scheduled, and would love for you to make it if the timezones work
> out.
> Best,
> Lucia
>
> --
> Lucia A. Perez
> School for Earth and Space Exploration
> PhD Candidate, Astrophysics
> <policy-on-using-previously-published-work.pdf>

Figure 5: Permission for Paper I (Chapter 2), from V. Tilvi.

ASU

Lucia Perez <laperez5@asu.edu>

IMPORTANT: Permission to put our LAE VPF paper in my dissertation?

Peter Laursen <pela@nbi.ku.dk>   Wed, Mar 23, 2022 at 11:29 AM
To: Lucia Perez <laperez5@asu.edu>

Hi Lucia,

That sounds great! And yes, of course you have my permission :)

I just had a look at the manuscript, and I think everything looks splendid. I'd very much like to join your dissertation, but yes, depending on what time you
defend it might be a bit late for me. Let me know when you know :)

All the best,
Peter

> On 23 Mar 2022, at 16.07, Lucia Perez <laperez5@asu.edu> wrote:
>
> Hi Peter!
> Hope you've been well! I'm FINALLY getting everything ready to defend and graduate! Currently scrambling to finish up some work, compile
> my dissertation, and schedule the defense. I have something to ask of you, and a few updates.
> On the VPF-LAGER paper: I've gotten the second-round referee response completely done and I'm quite sure it will be accepted, though I'm
> currently waiting for Sangeeta to settle some LAGER-specific collaboration questions before resubmitting. If you're curious, here's the
> Overleaf that will be sent in: https://www.overleaf.com/read/gpghkdtszzrj
> My big ask: Arizona State allows its PhD students to use to use published papers in dissertations as long as we have written permission
> (attached is the policy). Can I have your permission to put our manuscript in my dissertation as its own chapter? A response to this email is
> plenty.
> Thank you so much! I'll also share an invitation to my defense once it's scheduled, and would love for you to make it if the timezones work
> out.
> Best,
> Lucia
> --
> Lucia A. Perez
> School for Earth and Space Exploration
> PhD Candidate, Astrophysics
> <policy-on-using-previously-published-work.pdf>

Figure 6: Permission for Current Draft of Paper II (Chapter III), from P. Laursen.