The Astrophysical Journal, 977:119 (24pp), 2024 December 10 https://doi.org/10.3847/1538-4357/ad83cb
© 2024. The Author(s). Published by the American Astronomical Society.
OPEN ACCESS

ODIN: Identifying Protoclusters and Cosmic Filaments Traced by Lyα-emitting Galaxies

Vandana Ramakrishnan[1], Kyoung-Soo Lee[1], Maria Celeste Artale[2], Eric Gawiser[3], Yujin Yang[4], Changbom Park[5], Yi-Kuan Chiang[6], Robin Ciardullo[7,8], Arjun Dey[9], Caryl Gronwall[7,8], Lucia Guaita[2], Ho Seong Hwang[10,11], Sang Hyeok Im[10], Woong-Seob Jeong[4], Seongjae Kim[4], Ankit Kumar[2], Jaehyun Lee[4,5], Seong-Kook Lee[10,11], Byeongha Moon[4], Nelson Padilla[12], Alexandra Pope[13], Roxana Popescu[13], Akriti Singh[2,14], Hyunmi Song[15], Paulina Troncoso[16], Francisco Valdes[9], and Ann Zabludoff[17]

[1] Department of Physics and Astronomy, Purdue University, 525 Northwestern Avenue, West Lafayette, IN 47907, USA; ramakr18@purdue.edu
[2] Departamento de Ciencias Fisicas, Universidad Andres Bello, Fernandez Concha 700, Las Condes, Santiago, Chile
[3] Physics and Astronomy Department, Rutgers, The State University, Piscataway, NJ 08854, USA
[4] Korea Astronomy and Space Science Institute, 776 Daedeokdae-ro, Yuseong-gu, Daejeon 34055, Republic of Korea
[5] Korea Institute for Advanced Study, 85 Hoegi-ro, Dongdaemun-gu, Seoul 02455, Republic of Korea
[6] Institute of Astronomy and Astrophysics, Academia Sinica (ASIAA), Taipei 10617, Taiwan
[7] Department of Astronomy & Astrophysics, The Pennsylvania State University, University Park, PA 16802, USA
[8] Institute for Gravitation and the Cosmos, The Pennsylvania State University, University Park, PA 16802, USA
[9] NSF's National Optical-Infrared Astronomy Research Laboratory, 950 N. Cherry Avenue, Tucson, AZ 85719, USA
[10] Department of Physics and Astronomy, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea
[11] SNU Astronomy Research Center, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea
[12] Instituto de Astronomía Teórica y Experimental (IATE), CONICET-UNC, Laprida 854, X500BGR, Córdoba, Argentina
[13] Department of Astronomy, University of Massachusetts, Amherst, MA 01003, USA
[14] European Southern Observatory, Alonso de Córdova 3107, Vitacura, Santiago de Chile, Chile
[15] Department of Astronomy and Space Science, Chungnam National University, 99 Daehak-ro, Yuseong-gu, Daejeon, 34134, Republic of Korea
[16] Escuela de Ingeniería, Universidad Central de Chile, Avenida Francisco de Aguirre 0405, 171-0614 La Serena, Coquimbo, Chile
[17] Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721, USA
Received 2024 June 12; revised 2024 September 4; accepted 2024 October 5; published 2024 December 6

Abstract

To understand the formation and evolution of massive cosmic structures, studying them at high redshift, in the epoch when they formed the majority of their mass, is essential. The One-hundred-deg^2 DECam Imaging in Narrowbands (ODIN) survey is undertaking the widest-area narrowband program to date, to use Lyα-emitting galaxies (LAEs) to trace the large-scale structure (LSS) of the Universe on the scale of 10–100 cMpc at three cosmic epochs. In this work, we present results at $z = 3.1$ based on early ODIN data in the COSMOS field. We identify protoclusters and cosmic filaments using multiple methods and discuss their strengths and weaknesses. We then compare our observations against the IllustrisTNG suite of cosmological hydrodynamical simulations. The two are in excellent agreement, identifying a similar number and angular size of structures above a specified density threshold. We successfully recover the simulated protoclusters with $\log(M_{z=0}/M_\odot) \gtrsim 14.4$ in ∼60% of the cases. With these objects, we show that the descendant masses of our observed protoclusters can be estimated purely based on our 2D measurements, finding a median $z = 0$ mass of $\sim 10^{14.5} M_\odot$. The lack of information on the radial extent of each protocluster introduces a ∼0.4 dex uncertainty in its descendant mass. Finally, we show that the recovery of the cosmic web in the vicinity of protoclusters is both efficient and accurate. The similarity of our observations and the simulations implies that our structure selection is likewise robust and efficient, demonstrating that LAEs are reliable tracers of the LSS.

Unified Astronomy Thesaurus concepts: High-redshift galaxy clusters (2007); Large-scale structure of the universe (902); Lyα galaxies (978)

1. Introduction

According to the hierarchical theory of structure formation, matter is organized into a *cosmic web*, comprised of linear filaments intersecting at nodes of high density surrounded by vast voids (J. R. Bond et al. 1996; V. Springel et al. 2005; M. Boylan-Kolchin et al. 2009). This large-scale structure (LSS) determines how much cold gas is available to a galaxy and the likelihood of a merger or interaction with another galaxy, thereby acting as one of the fundamental drivers of galaxy evolution.

Out to $z \approx 1.5$, redshift surveys and other observational techniques have enabled the selection of samples of galaxies inhabiting clusters, groups, and filaments (e.g., P. R. M. Eisenhardt et al. 2008; Y. Koyama et al. 2014; E. S. Rykoff et al. 2014, 2016; E. Tempel et al. 2014; L. E. Bleem et al. 2015; N. Malavasi et al. 2017; M. Hayashi et al. 2018; A. H. Gonzalez et al. 2019). These studies show that galaxies in cluster or group environments tend to be older and more massive, and are more likely to have ceased star formation than those in the field (e.g., Y.-j. Peng et al. 2010; R. F. Quadri et al. 2012; R. F. J. van der Burg et al. 2013). Filaments may have a weaker but similar effect and may be responsible for preprocessing galaxies that are falling into cluster or group environments (e.g., F. Sarron et al. 2019; J. M. Salerno et al. 2019).

At Cosmic Noon ($z \gtrsim 2$), when the global star formation rate reached its peak (P. Madau & M. Dickinson 2014), these

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Table 1
Existing Samples of $z = 3.1$ LAEs

Sample	No. of Fields	Total Area (deg^2)	Depth ($erg\ s^{-1}$)	EW_0 (Å)	Δz	No. of LAEs	Surface Density ($arcmin^{-2}$)
This work	1	9.0	1.9×10^{42}	20	0.059	5691	0.21
R. Ciardullo et al. (2002)	1	0.13	4.5×10^{42}	20	0.045	9	0.02
T. Hayashino et al. (2004)	1	0.21	4.0×10^{42}	38	0.063	283	0.37
C. Gronwall et al. (2007)	1	0.28	1.3×10^{42}	20	0.024	259	0.26
M. Ouchi et al. (2008)	1	0.98	1×10^{42}	64	0.061	356	0.10
Y. Matsuda et al. (2009)	1	0.20	1.3×10^{42}	20	0.061	127	0.18
R. Ciardullo et al. (2012)	1	0.28	2.4×10^{42}	20	0.047	199	0.20
T. Yamada et al. (2012)	4	2.42	1.5×10^{42}	46	0.063	2161	0.25
R. M. Bielby et al. (2016)	5	1.25	1.0×10^{42}	65	0.063	643	0.14
J. Matthee et al. (2017)	1	0.85	1.0×10^{43}	12	0.082	65	0.02

environmental effects are predicted to be even more dramatic. In the high-density regions within the LSS, the accretion rates of infalling gas and the incidence of galaxy interactions are expected to be greatest (see discussion in B. C. Lemaux et al. 2022, and references therein), fostering both enhanced in situ star formation and black hole activity. A popular hypothesis is that highly dissipative gas-rich mergers help the efficient feeding of gas into the central black hole and trigger active galactic nuclei (AGNs), which may ultimately quench the star formation activity (e.g., P. F. Hopkins et al. 2006). These expectations are indeed in line with the heightened SF and AGN activities found in a handful of *protocluster* systems (e.g., C. M. Casey et al. 2015; H. Umehata et al. 2015; T. Wang et al. 2016; I. Oteo et al. 2018; Y. Harikane et al. 2019; B. C. Lemaux et al. 2022) as well as the emergence of quenched galaxies in such environments (e.g., K. Shi et al. 2021; K. Ito et al. 2023).

Yet, the role that the LSS environment plays in galaxy formation at Cosmic Noon remains underexplored. Our limited knowledge is due to a combination of factors, including the observational limitations of measuring precise redshifts of faint, high-redshift galaxies, the inherent scarcity of massive cosmic structures, and our incomplete grasp of the indicators for the locations of these structures. The lack of readily identifiable signatures—such as a hot intracluster medium and/or a concentration of quiescent galaxies—in young, yet-to-be-virialized structures of mostly star-forming galaxies leads to a strong reliance on spectroscopy for finding protoclusters. While the lack of large, statistical samples prevent us from disentangling the effects of cosmic variance from general properties of protocluster galaxies, studies from heterogeneously selected samples can lead to seemingly conflicting conclusions. Although cosmic filaments connected to these protoclusters likely play a vital role in replenishing fresh gas for sustained star formation, such medium-density features are even more difficult to discern than dense protocluster cores.

The One-hundred-deg^2 DECam Imaging in Narrowbands (ODIN; K.-S. Lee et al. 2024) survey is designed to obtain large and *uniformly selected* samples of protoclusters and filaments at three cosmic epochs ($z = 2.4$, 3.1, and 4.5) using Lyα-emitting galaxies (LAEs) as tracers of underlying matter distribution. As the most common electron transition in the universe, Lyα emission traces ionized and/or excited gas from star formation, black hole activity, and the gravitational collapse of dark matter halos. A large fraction of low-luminosity star-forming galaxies (which dominate the cosmic star formation rate density; N. A. Reddy & C. C. Steidel 2009) show Lyα emission (D. P. Stark et al. 2010). These LAEs tend to have lower stellar masses, younger population ages, and less internal extinction than systems selected via their broadband colors, and low galaxy bias (E. Gawiser et al. 2007; H. Kusakabe et al. 2018; L. H. Weiss et al. 2021; M. White et al. 2024). These traits make LAEs a most efficient tracer of the underlying dark matter distribution, one that can be used to constrain cosmology (e.g., K. Gebhardt et al. 2021; M. White et al. 2024) and the most massive cosmic structures (A. Dey et al. 2016; Y. Huang et al. 2022). Upon completion, ≈600 protoclusters are expected to be discovered by ODIN, which will facilitate robust statistical investigations of the cosmic evolution of protoclusters and their galaxy inhabitants.

In this paper, we use the early ODIN data in the COSMOS field to present our selection of protoclusters and cosmic filaments. Building on the results presented in V. Ramakrishnan et al. (2023), we calibrate and fine-tune our LSS detection methods by carrying out careful comparisons with cosmological hydrodynamical simulations. The outline of this paper is as follows. In Section 2, we give details of the observational and simulation data. In Sections 3 and 4, we describe how we construct LAE surface density maps and how we use them to detect protoclusters and filaments. We validate our procedures and interpret the results using the mock data created from the simulations in Section 5. Finally, the properties of our observationally selected structures are discussed in Section 6 followed by a summary of our findings in Section 7. Throughout this paper, we assume a Planck cosmology (Planck Collaboration et al. 2016): $\Omega_\Lambda = 0.6911$, $\Omega_b = 0.0486$, $\Omega_m = 0.3089$, $H_0 = 100\,h\,km\,s^{-1}\,Mpc^{-1}$ and $h = 0.6774$. Distances are given in units of comoving Mpc (cMpc) unless noted otherwise.

2. Observational and Simulation Data

2.1. The ODIN Survey

The ODIN survey is conducting the widest-area deep narrowband imaging program to date, using three custom narrowband filters (*N*419, *N*501, and *N*673) to identify redshifted Lyα emission. In this work, we make use of the year 1 ODIN data taken with the *N*501 filter ($\lambda_C/\Delta\lambda = 5014/75$ Å; $\bar{z}/\Delta z = 3.12/0.06$) in the extended COSMOS field. The data have a resolution of $0\farcs27\,pix^{-1}$ and a 5σ magnitude limit of 25.4 AB over a total area of $\sim 9\,deg^2$, allowing us to detect Lyα emission down to a line flux of $\sim 2.1 \times 10^{-17}\,erg\,s^{-1}\,cm^{-2}$. As shown in Table 1, our

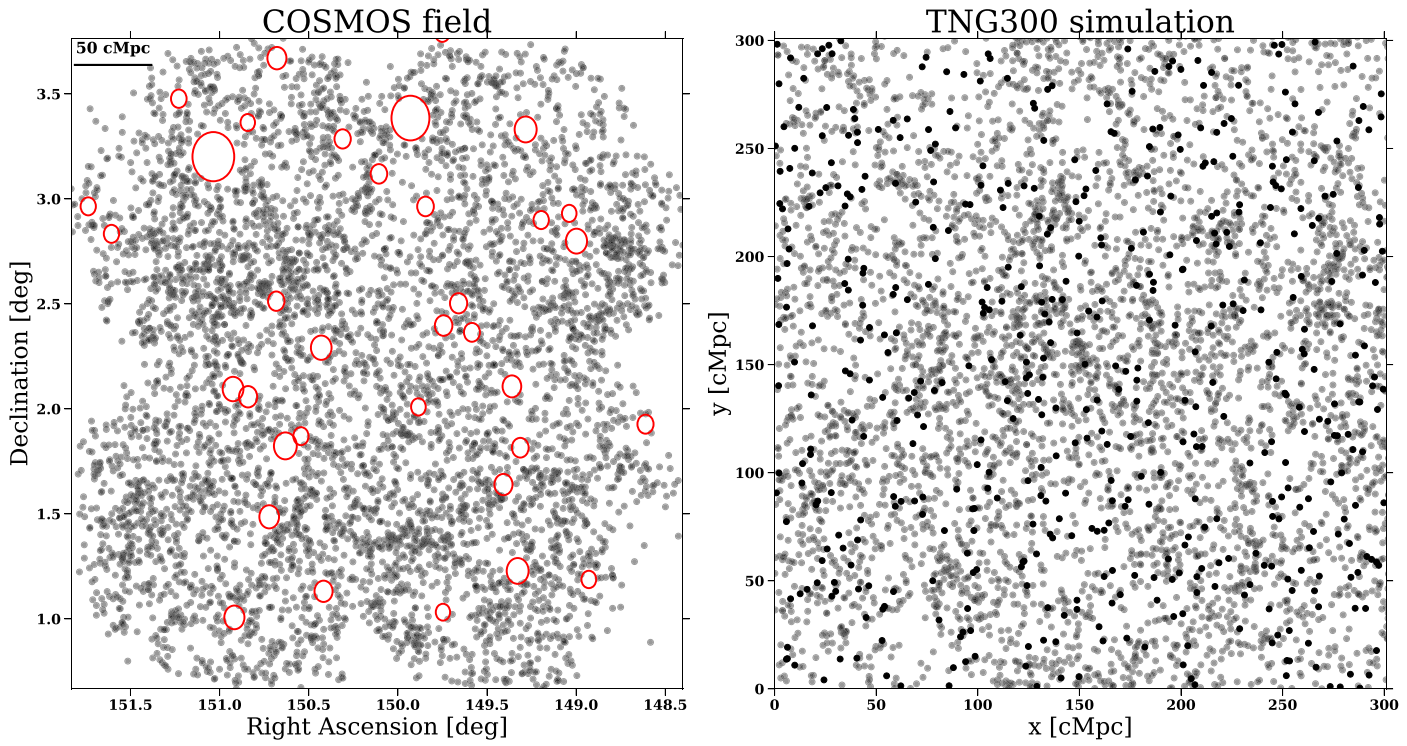

Figure 1. Positions of the observed ODIN LAEs (left panel, see text in Section 2.1) and the mock LAEs from a randomly chosen slice of TNG300 (right panel, see text in Section 2.2). In the left panel, the holes left by bright stars are shown by red circles.

narrowband data are similar in depth to the deepest surveys undertaken at similar redshifts while being ∼4 times or more larger in area. For more details about the survey fields and observing strategy, we refer interested readers to K.-S. Lee et al. (2024).

The positions of the ODIN LAEs are shown in the left panel of Figure 1. The method for selecting these LAEs is detailed in N. M. Firestone et al. (2024). Briefly, we identify LAEs as *N*501-detected sources exhibiting a narrowband excess over the continuum, corresponding to a rest-frame equivalent width of 20 Å. The continuum magnitude is calculated as a weighted combination of the magnitude in two broadband filters (*g* and *r* for *N*501 LAEs). We reject sources flagged for saturated pixels or other image defects and those close to bright stars, with the total excluded area being $\sim$1.5 deg^2. The resultant gaps left in the LAE source distribution are highlighted in Figure 1. The largest of these is $\sim$144 arcmin^2 in area, but there are only two such holes in the field. The remainder are much smaller ($\lesssim$36 arcmin^2). The continuum is estimated using broadband data from the Hyper Suprime-Cam Subaru Strategic Program (HSC-SSP; H. Aihara et al. 2018a, 2018b) second data release (H. Aihara et al. 2019).

The final LAE sample comprises 5691 sources over $\approx$7.5 deg^2, corresponding to a surface density of 0.21 arcmin^{-2}. This is comparable to other narrowband surveys with similar filter widths and narrowband depths (see Table 1). Spectroscopic follow-up efforts for ODIN LAEs are ongoing, but sufficient spectra have been taken to obtain a preliminary estimate of the purity of our LAE sample. Of the sources that yielded a redshift, $\approx$97% are confirmed as LAEs (V. Ramakrishnan et al. 2024, in preparation).

2.2. Building Mock ODIN Observations with TNG

To build a concrete framework in which we can interpret our observations, we use the IllustrisTNG300-1 simulation (hereafter TNG300; A. Pillepich et al. 2018a, 2018b; D. Nelson et al. 2019) and define our mock LAE samples. TNG300 provides the largest volume (302.6 cMpc on a side) of all available simulations of the IllustrisTNG suite. Given the rarity of massive galaxy (proto)clusters, this is especially crucial for our work. All TNG simulations assume the Planck cosmology (Planck Collaboration et al. 2016). The surface area viewed along the *X*-, *Y*-, or *Z*-direction is $\approx$90,000 cMpc^2, well matched to the angular extent of the ODIN data in the COSMOS field, $\approx$120,000 cMpc^2, and the simulation box is several times larger than the ODIN radial extent of $\sim$60 cMpc at $z = 3.1$. TNG300 has a baryon mass resolution of $1.1 \times 10^7 M_\odot$ and a dark matter mass resolution of $5.9 \times 10^7\, M_\odot$. In our analysis, we employ galaxies with stellar mass greater than $10^7\, M_\odot$, corresponding to a halo mass of $\gtrsim 10^9\, M_\odot$; as we do not use any stellar or gas physics in our analysis, the resolution is sufficient for our purposes.

In constructing mock LAE samples, our primary goal is to mimic the spatial distribution of the ODIN LAEs as closely as possible so that we can evaluate their utility as tracers of the LSS. Understanding the complex behavior of Lyα radiative transfer requires vastly higher resolution simulations and thus is outside the scope of this work. Our selection of mock LAEs from among the TNG galaxies is hence based only on their stellar mass; we do not attempt a selection based on the star formation rate or modeled Lyα flux or equivalent width as in, e.g., M. Dijkstra & E. Westra (2010) and J. Ravi et al. (2024). We describe our procedure in detail below.

First, we match the 75 Å (60 cMpc at $z = 3.1$) full width at half-maximum of $N501$ by creating cosmic "slices" from the TNG $z = 3$ snapshot with ≈80 cMpc in line-of-sight thickness. The filter and thus the window function are not a perfect top hat in shape. To emulate this effect, we assign the LAE selection probability to match the shape of the filter transmission function. In all things being equal, the probability of being selected as a mock LAE is 1, 0.5, and 0 at the distance of 0, 30, and 40 cMpc from the center of the slice, respectively. In practice, the probability of an LAE being detected at a given position along the redshift direction also depends on its line luminosity, as bright LAEs are more likely to be detected when they fall on the wings of the filter than fainter ones, and hence are detected over a larger volume. However, since we are interested only in the average number density of *all* LAEs, irrespective of line luminosity, this does not significantly affect our analysis.

Second, we aim to reproduce the small- and large-scale clustering of LAEs. The galaxy bias for $z \sim 3$ LAEs is relatively low at $b \lesssim 2$ (E. Gawiser et al. 2007; M. Ouchi et al. 2008; M. White et al. 2024), suggesting that LAEs have low stellar mass content and are hosted by low-mass halos (e.g., L. Guaita et al. 2010; H. Kusakabe et al. 2018). To match the small-scale clustering, we must simultaneously match the LAE overdensity distribution across the field; the details of this measurement are presented in Section 3. Motivated by the findings of A. Hagen et al. (2014), we select the mock LAEs such that their stellar masses obey a lognormal distribution: i.e., $\log M_*/M_\odot$ is a Gaussian function, with mean and standard deviation (μ, σ). Lowering μ values would shift the host halos to lower masses (thus lower bias). A larger scatter σ (while fixing μ) would permit both higher- and lower-mass halos to host LAEs, thereby changing how LAEs trace these halos. We find that $\mu = 8.75$ and $\sigma = 0.75$ yield the best fit to our surface density measurements. This choice nearly perfectly reproduces the observed distribution of LAE surface density values, as we show in Section 5.1. Our best-fit parameters are fully consistent with the stellar mass distribution of LAEs based on reported by A. Hagen et al. (2014) by fitting their spectral energy distributions, $\log (M_*/M_\odot) = 8.97^{+0.60}_{-0.71}$. They are also similar to the result of C. J. Vargas et al. (2014), who find $\log (M_*/M_\odot) = 8.45^{+0.72}_{-0.67}$ for LAEs at $z = 2.1$.

Finally, we match the measured sky density of LAEs. The $N501$ LAE surface density is 0.21 arcmin^{-2} (see Section 2.1). In comparison, the number of TNG galaxies selected based on the above criteria is typically 10 times greater. Additionally, we assume 10% of our LAEs may be contaminants. This number is based on the fraction of our spectroscopic targets in the two faintest bins ($N501 = 24.5$–25.0 and 25.0–25.5 AB) that yielded a redshift, which is 94.3% and 85.9%, respectively (V. Ramakrishnan et al. 2024, in preparation). By doing so, we are assuming that all sources that did not result in a redshift are interlopers. Our spectroscopic success rate (i.e., the fraction of sources that yield a redshift for which it is within the expected range) is 97%, so our estimate is conservative. Changing it to 5% does not change our results. From the selected TNG galaxies, we randomly choose a subset whose number is equal to 90% of the LAE density; assuming that LAE sample contaminants are unclustered, the remaining 10% is drawn as a purely random distribution within the same region.

To compare directly with our observations, we utilize the progenitors of 30 of the most massive galaxy clusters in the final simulation box of TNG300, ranked based on their total stellar mass. The stellar mass, halo mass, and group ID of these clusters at $z = 0$ are found in M. Andrews et al. (2024). We identify the protoclusters by tracing the main progenitor branch of the merger trees of the $z = 0$ clusters to their $z = 3$ predecessors. The descendant masses of the protoclusters range from $1.5 \times 10^{15} M_\odot$ for the most massive system to $2 \times 10^{14} M_\odot$ for the least massive. We generate cosmic slices, aligning the midpoint of each with that of a protocluster using the periodic boundary conditions of the simulation. Slicing the TNG volume along the X-, Y-, and Z-axis produces 90 slices for comparison. In a given slice, typically, a few other protoclusters from the top 30 most massive sample are included although, depending on their positions in the redshift direction, they may be only partially represented. The positions of the mock LAEs in a randomly chosen slice are shown in the right panel of Figure 1.

3. Tracing Large-scale Structure

This section aims to delineate the LSS in our observations by constructing the LAE surface density maps. We employ two methods: first, smoothing over the LAE positions with a fixed Gaussian kernel; and second, constructing the Voronoi diagram of the LAEs. These two methods are summarized below but are discussed in more detail in V. Ramakrishnan et al. (2023).

3.1. Gaussian Smoothing

The simplest approach to measure the surface density across a field involves smoothing over the positions of the galaxies within it using a fixed-size kernel. We employ a 2D Gaussian kernel with FWHM 10 cMpc. This FWHM is decided following the methodology of T. Bădescu et al. (2017), also utilized in V. Ramakrishnan et al. (2023). The kernel size is chosen such that the resultant surface density map maximizes the total probability at the positions of the LAEs in the real data. This is achieved through a leave-one-out cross-validation, where the likelihood of finding a point at the location $\boldsymbol{r}_j$ of the jth data point is estimated as

$$p(\boldsymbol{r}_j) = \sum_{i \neq j} \frac{1}{\sqrt{2\pi}\,\sigma} \exp \frac{-(\boldsymbol{r}_i - \boldsymbol{r}_j)^2}{2\sigma^2}. \quad (1)$$

The optimum σ value is the one that maximizes $\prod_j p(\boldsymbol{r}_j)$.

Before creating the LAE surface density map, we fill in the holes left by removing the sources near bright stars and image defects with uniformly distributed random points with surface density matched to that of the LAEs. We do not find significant differences between the surface density maps produced before and after carrying out this procedure. After convolving with the kernel, the overdensity map is computed by dividing the $\Sigma_{\rm LAE}$ map by the mean surface density, $\overline{\Sigma}_{\rm LAE}$:

$$(1 + \delta_{\rm LAE}) = \frac{\Sigma_{\rm LAE}}{\overline{\Sigma}_{\rm LAE}}. \quad (2)$$

The mean and standard deviation are determined by fitting the LAE surface density distribution with a Gaussian function, $\exp[-(\Sigma_{\rm LAE} - \overline{\Sigma}_{\rm LAE})^2/2\sigma^2]$. The fit is restricted to within $\pm 1.5\sigma$ (after iterative σ-clipping) to ensure that our estimate is not biased by the presence of multiple high LAE overdensities, which show up at the high $(1 + \delta_{\rm LAE})$ end of the distribution.

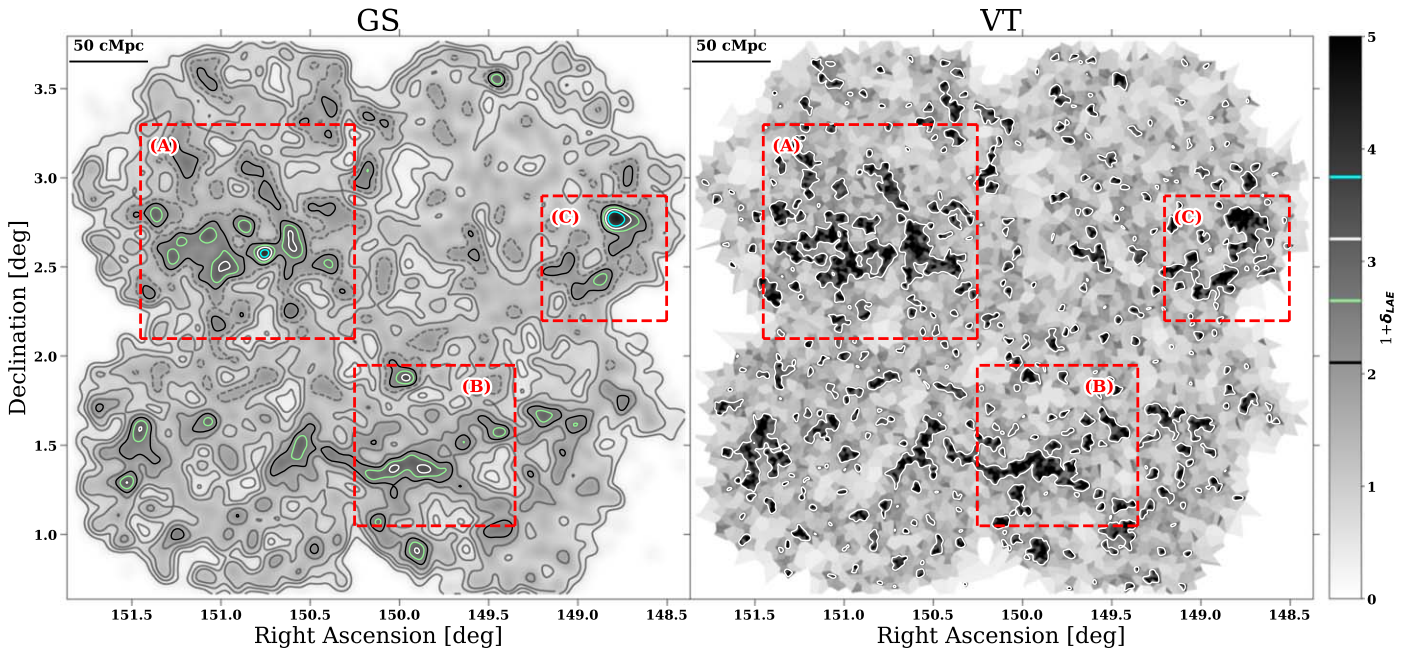

Figure 2. Left: Gaussian smoothed (GS) LAE surface density map. Black, green, white, and cyan contours indicate LAE overdensities of 2σ, 3σ, 4σ, and 5σ significance respectively; the corresponding values of $(1+\delta_{\rm LAE})$ are marked on the color bar. The three largest overdensity complexes are highlighted in red. Right: VT-based density map. White contours indicate overdensities of 3σ significance. The LSS revealed by the surface density maps is highly clumpy and irregular. Nonspherical/irregular overdensities are better delineated at fixed significance in the VT map than in the GS map. The clover shape of the surface density maps arises from the HSC-SSP data, which consists of four Deep and one central UltraDeep pointing in the COSMOS field.

The left panel of Figure 2 shows the resultant $(1+\delta_{\rm LAE})$ map, which we will refer to as the Gaussian smoothed (GS) map, hereafter. It shows overdense regions that are both strongly clustered and highly irregular in shape, consistent with expectations from the hierarchical theory of structure formation, in which smaller structures continuously merge to form larger ones, a phenomenon supported by hydrodynamical simulations (e.g., M. Boylan-Kolchin et al. 2009). The three most prominent overdensity complexes, highlighted by red boxes and labeled as complexes A, B, and C, were discussed in detail in V. Ramakrishnan et al. (2023).

Smoothing with a fixed kernel is most effective at identifying structures with a size and shape similar to that of the kernel itself. Thus, the Gaussian smoothing method may not adequately capture nonisotropic features, potentially resulting in an underestimation of the significance of many observed structures. To address this, we explore tessellation-based methods in Section 3.2.

3.2. Voronoi Tessellation

Tessellation-based density estimates offer the advantage of being scale-independent and do not assume any specific shape or size for the underlying structures. The two algorithms most commonly used in the literature are the Voronoi tessellation (VT; e.g., A. Dey et al. 2016; O. Cucciati et al. 2018; B. C. Lemaux et al. 2018; D. Hung et al. 2020) and the Delaunay tessellation (e.g., N. Malavasi et al. 2021; F. Sarron & C. J. Conselice 2021). B. Darvish et al. (2015) find, through analysis of a simulated data set, that the Delaunay tessellation fares more poorly at estimating the "true" surface density value at a given point as compared to the VT. Additionally, the method tends to overestimate the surface density in overdense regions. Considering these results, we opt for the VT algorithm as outlined below.

In a 2D case, the VT divides a plane into distinct cells based on the positions of a set of generating points, e.g., the sky locations of LAEs. The Voronoi cell of each generating point includes all regions in the plane that are closer to it than to any other generating point. Consequently, the area of each cell is a measure of surface density: cells associated with LAEs in overdense regions will be small due to the proximity of numerous other LAEs, whereas those in underdense regions will be larger. Since each cell contains a single LAE, its LAE surface density Σ_i is given by

$$\Sigma_i = \frac{1}{A_i} \tag{3}$$

where A_i is the area of the ith cell. As before, the masked regions are filled in before constructing the Voronoi diagram.

In Figure 2, we show the VT $(1+\delta_{\rm LAE})$ map together with the GS map. While both maps detect the most significant structures, the VT method detects them at higher significance. As expected, the method also identifies a greater number of overdensities by capturing irregular or anisotropic features.

4. Features of the LSS in Observations: Protoclusters and Filaments

Galaxy protoclusters represent some of the most striking features of the LSS at high redshift and are expected to be observed as significant galaxy overdensities spanning $\approx$10 cMpc in scale (Y.-K. Chiang et al. 2013). According to the hierarchical theory of structure formation, massive halos hosting protoclusters are connected to filaments of the cosmic web (e.g., U. Kuchner et al. 2022) along which pristine gas is being accreted to feed star formation. Indeed, these predictions align qualitatively with the features observed in our $(1+\delta_{\rm LAE})$ map. Multiple overdensities with angular scales of 5–10 cMpc are evident in Figure 2 (right), frequently clustered together to

form *complexes* comprising two to five adjacent overdensities. Additionally, regions of high overdensity, where $(1+\delta_{\rm LAE}) > 3$, are interconnected by more moderate "bridges" with $(1+\delta_{\rm LAE}) = 2$–3. Several features show a distinctly linear morphology, reminiscent of cosmic filaments (for example, the northwestern corner of "complex A" and the southeastern corner of "complex C," see V. Ramakrishnan et al. 2023, for more detail).

Motivated by these observations, our next objective is to pinpoint the locations of protoclusters and filaments of the cosmic web using LAEs as tracers. By directly comparing our observations with predictions based on the TNG300 mock LAEs, we will also test how robust our protocluster candidates are and measure their key properties.

4.1. Selecting Protoclusters from Density Maps

Protoclusters are generally defined as regions that will evolve into virialized structures with masses $\gtrsim 10^{14} M_{\odot}$ by $z = 0$ (e.g., R. A. Overzier 2016). In distant look-back times, protoclusters are loosely bound regions comprised of multiple dark matter halos. This poses a challenge in measuring their physical extent. In this study, we define a protocluster as a region exhibiting significant size and overdensity, unlikely to occur by chance alone, and containing sufficient mass that could plausibly collapse to form a single halo with mass greater than or equal to $10^{14} M_{\odot}$ by $z = 0$.

According to this operational definition, selecting protoclusters entails identifying contiguous regions above a preset overdensity threshold and minimum area. This procedure mirrors source detection in pixelated astrophysical images. Therefore, adopting approaches similar to those outlined by B. C. Lemaux et al. (2018), D. Hung et al. (2020), and F. Sarron & C. J. Conselice (2021), we utilize source detection software on our density maps for protocluster selection.

We begin by pixelating the surface density maps by interpolating them over a 2D grid of positions with pixel size $3\farcs6$. Our grid size, corresponding to 115 ckpc at $z = 3.1$, is small enough to clearly delineate the boundaries of the structures, as it is 2 orders of magnitude smaller than the typical size of a protocluster. For source detection, we use SEP (K. Barbary 2016), a Python implementation of the SExtractor (E. Bertin & S. Arnouts 1996) software. The number of detected structures depends strongly on the detection threshold (`DETECT_THRESH`) and minimum area (`DETECT_MINAREA`). Our goal is to optimize these parameters to maximize the identification of robust candidates while minimizing the inclusion of spurious objects resulting from chance alignments of LAEs.

To assess contamination, we randomly select a subset of *all* continuum and line-emitting sources detected within the *N*501 image, matching the number and narrowband magnitude distribution of LAEs in the field. We then generate GS and VT maps for these random points following the same procedure as for the LAE surface density maps. Since these points are distributed across a wide range of redshift, any overdensity observed in these "random maps" is unlikely to be genuine but rather the result of chance alignments. Hence, the contamination fraction for a specific set of detection parameters can be approximated by dividing the number of sources detected in the random map by the number detected in the LAE surface density map. This evaluation is depicted in the top and bottom right panels of Figure 3 for the VT and GS map, respectively, with the contamination fraction averaged over 30 iterations.

For the GS-selected protoclusters, the contamination fraction is low for all sets of detection parameters used. Our fiducial setup, `DETECT_THRESH` of 3σ and `DETECT_MINAREA` of 4600 pixels ($\simeq$60 cMpc^2), yields a contamination fraction of $<0.1\%$ where σ is the (σ-clipped) standard deviation of $\delta_{\rm LAE}$. Compared to the GS map, the pixel-to-pixel density fluctuations are much greater in the VT map, introducing much higher noise in structure detection. This likely indicates that the scale at which the statistics are being sampled is too small. We mitigate this effect by smoothing the map with a 5 cMpc FWHM Gaussian kernel. On average, this kernel interpolates over $\sim$4 LAEs. It is thus large enough to reduce the noise but small enough to avoid adjacent structures blending into one considering that a typical diameter of a protocluster at $z \sim 3$ is 10 cMpc. We choose `DETECT_THRESH` of 4.5σ and `DETECT_MINAREA` of 3000 pixels ($\simeq$40 cMpc^2), yielding the contamination rate of 20%.

Figure 4 (top and middle rows) shows the protoclusters identified from the GS and VT-based maps within the highlighted complexes A, B, and C. The GS map predominantly captures the largest structures, with smaller and more irregular formations remaining undetected; nonetheless, the identified structures exhibit high robustness. In contrast, the VT map reveals numerous structures overlooked by the GS map but is also subject to higher noise levels and is more susceptible to chance alignments of LAEs. The detailed comparison of different detection methods is presented in Section 6.

4.2. Selecting Protoclusters with HDBSCAN

Density-based clustering techniques, as a subset of unsupervised machine learning algorithms (H.-P. Kriegel et al. 2011), discern clusters by identifying regions of high point density in data space, set apart by areas of low density. These methods can be applied to pinpoint protoclusters from the clustering of the LAEs. Here, we employ the hierarchical density-based spatial clustering of applications with noise (HDBSCAN) algorithm (R. J. G. B. Campello et al. 2013). HDBSCAN offers the advantage of detecting clusters with varying densities, regardless of their shape, provided they exceed a specified size threshold. Briefly, this algorithm measures density at each location, finding clusters as peaks in the density distribution separated by troughs. Density estimation is based on the distance to the *K*th nearest neighbor.

In HDBSCAN, each point is classified into *clusters* or *noise* based on two parameters: the *K*th nearest neighbor distance and the minimum cluster size. The latter dictates the minimum number of data points constituting a cluster. In the Python implementation of the HDBSCAN algorithm (`hdbscan`; L. McInnes et al. 2017), these parameters are denoted as `min_samples` and `min_cluster_size`, respectively. In our setup, we opt for values of 15 and 10, respectively (i.e., a minimum cluster size of 15 LAEs and the 10th nearest neighbor distance; see Appendix A). Additionally, we require that each cluster exhibits a median surface density above a specified threshold. This surface density is calculated as the number of LAEs divided by the enclosed area, computed as the sum of the areas of all Voronoi cells encompassing the cluster members. We set the threshold at 0.23 arcmin^{-2} (i.e., $\gtrsim 1\sigma$ above the field mean calculated from the VT map), which yields a

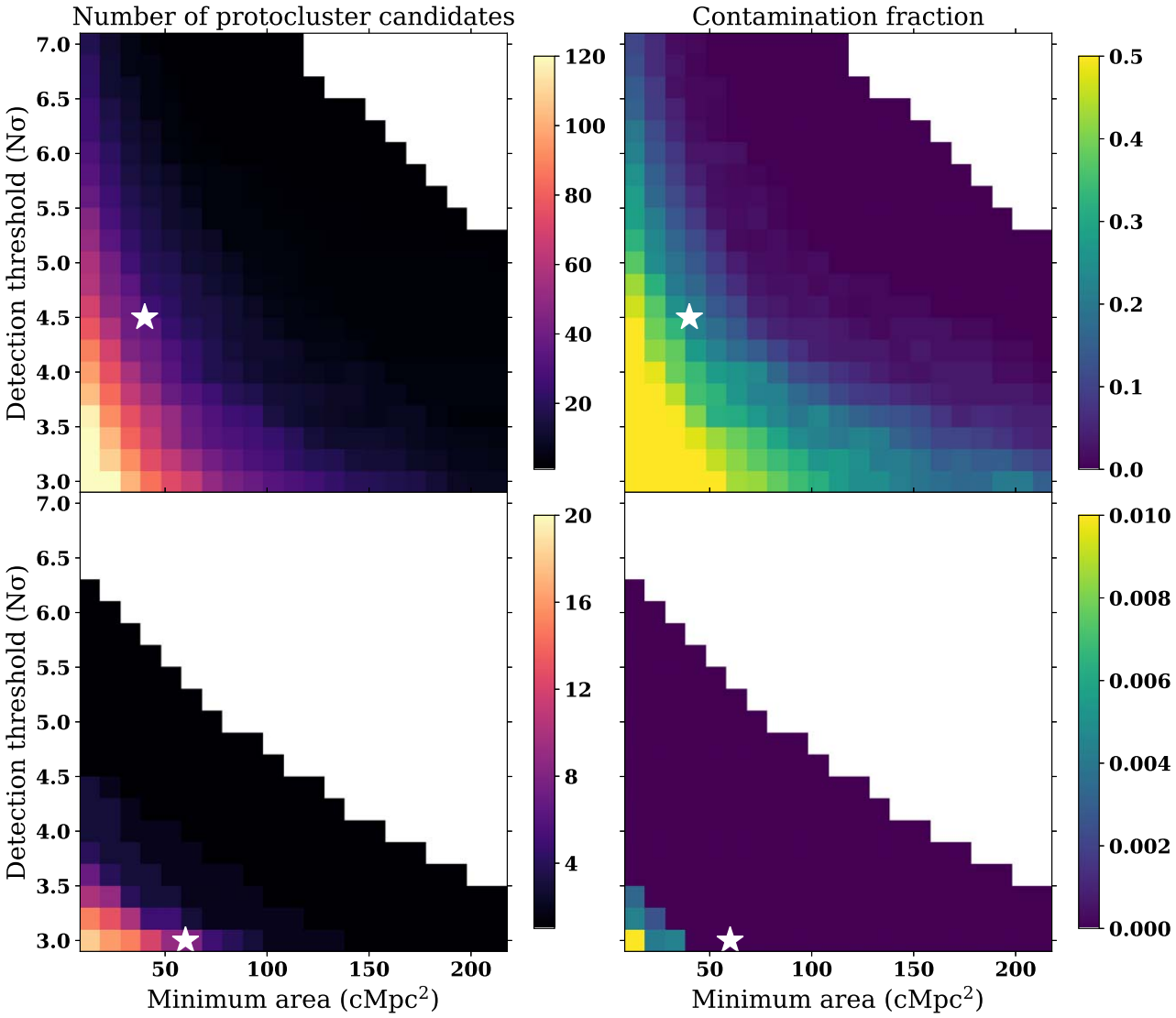

Figure 3. Number of candidates and contamination fraction for protoclusters selected from the VT map (top) and GS map (bottom), for different density thresholds and minimum areas. The density threshold is expressed in units of Nσ above the field surface density. White stars indicate the detection settings adopted for this work. While the VT map suffers from higher contamination than the GS map, it also detects far more candidates.

contamination rate of ≈20% similar to that for the VT sample. The HDBSCAN-detected protoclusters are illustrated in the bottom row of Figure 4. Symbols of the same color denote membership within the same structure, while gray circles represent LAEs outside of protoclusters. They are also shown in Figure 5.

4.3. Cosmic Filaments with DisPerSE

The cosmic web at low redshift ($z \lesssim 1$) has been investigated in detail through multiple surveys (e.g., M. Alpaslan et al. 2014; E. Tempel et al. 2014; N. Malavasi et al. 2017). These studies have yielded valuable insight into the complex interplay of filaments and the clusters at the nodes of the cosmic web. Notably, F. Sarron et al. (2019) and J. M. Salerno et al. (2019) found evidence that galaxies infalling into clusters along filaments have a higher passive fraction than those being accreted from other directions. Moreover, several studies observed that, on scales of 5–10 cMpc, massive and quiescent galaxies are preferentially located closer to filaments than low-mass and star-forming galaxies (K. Kraljic et al. 2017; N. Malavasi et al. 2017; C. Laigle et al. 2018). This indicates that filaments may be preprocessing the galaxies being accreted along them. At higher redshift, however, the role of the cosmic web on galaxy formation remains largely observationally unconstrained. Motivated by the discernible presence of filamentary structures in the LAE surface density map, we aim to detect cosmic web filaments in our observations in this section. We explore the reliability of our filament detection in Section 5.3.

To identify cosmic filaments, we use the Discrete Persistent Structure Extractor (DisPerSE; T. Sousbie 2011; T. Sousbie et al. 2011), a widely utilized tool in both observational (e.g., N. Malavasi et al. 2017, 2020; F. Sarron et al. 2019) and simulation studies (e.g., C. Byrohl & D. Nelson 2023; S. H. Im et al. 2024). DisPerSE utilizes the Delaunay Tessellation Field Estimator (W. E. Schaap & R. van de Weygaert 2000; R. van de Weygaert & W. Schaap 2009; M. C. Cautun & R. van de Weygaert 2011) to create a discrete representation of the density field based on a given set of points. It then locates

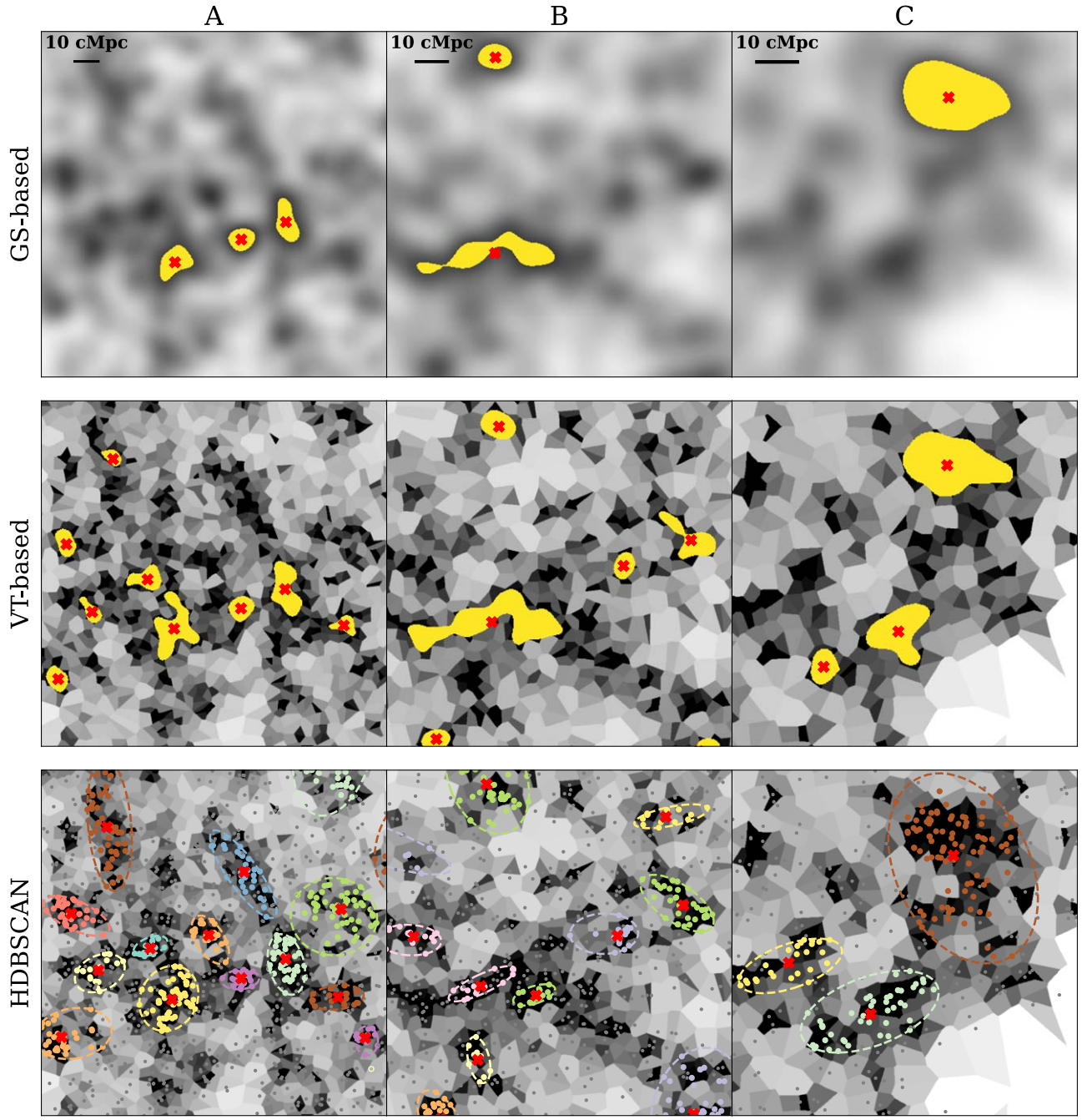

Figure 4. Protocluster candidates selected from the GS map (top row), the VT map (middle), and with HDBSCAN (bottom). Red crosses indicate the geometric centers of the selected overdensities. Yellow swathes indicate the extent of the protoclusters in the upper two rows. In the bottom row, colored points indicate LAEs classified as belonging to a cluster by HDBSCAN, enclosed by an ellipse for clarity, while gray points are classified as noise.

critical points within this density field approximation and defines filaments as arcs connecting saddle points to maxima. The filaments are constructed by creating short segments tangent to the gradient of the density field at each point. To account for data noise, DisPerSE employs the concept of persistence, which measures the density contrast between critical points defining a filament. Persistence represents the range of density thresholds over which a filament connecting two critical points remains significant relative to the noise. This significance is expressed in units of σ, where an $N\sigma$ filament corresponds to a probability under a 1D Gaussian distribution. We stress that this method of measuring the persistence level does not imply the use of a Gaussian distribution within DisPerSE but rather is purely an expression of the likelihood of the filaments arising from noise. For example, a persistence level of 3σ (2σ) means that the extracted filament has a 99.7% (95.5%) probability of being a true feature.

After filling in the voids left by star masks, we run DisPerSE on the positions of the LAEs with a persistence of 2σ. We note that there is no significant difference between the filaments recovered with and without filling in the voids, as shown in V. Ramakrishnan et al. (2023). The resultant filament network is illustrated in the bottom right panel of Figure 5 as red curves superimposed on the VT density map in gray scale. The figure

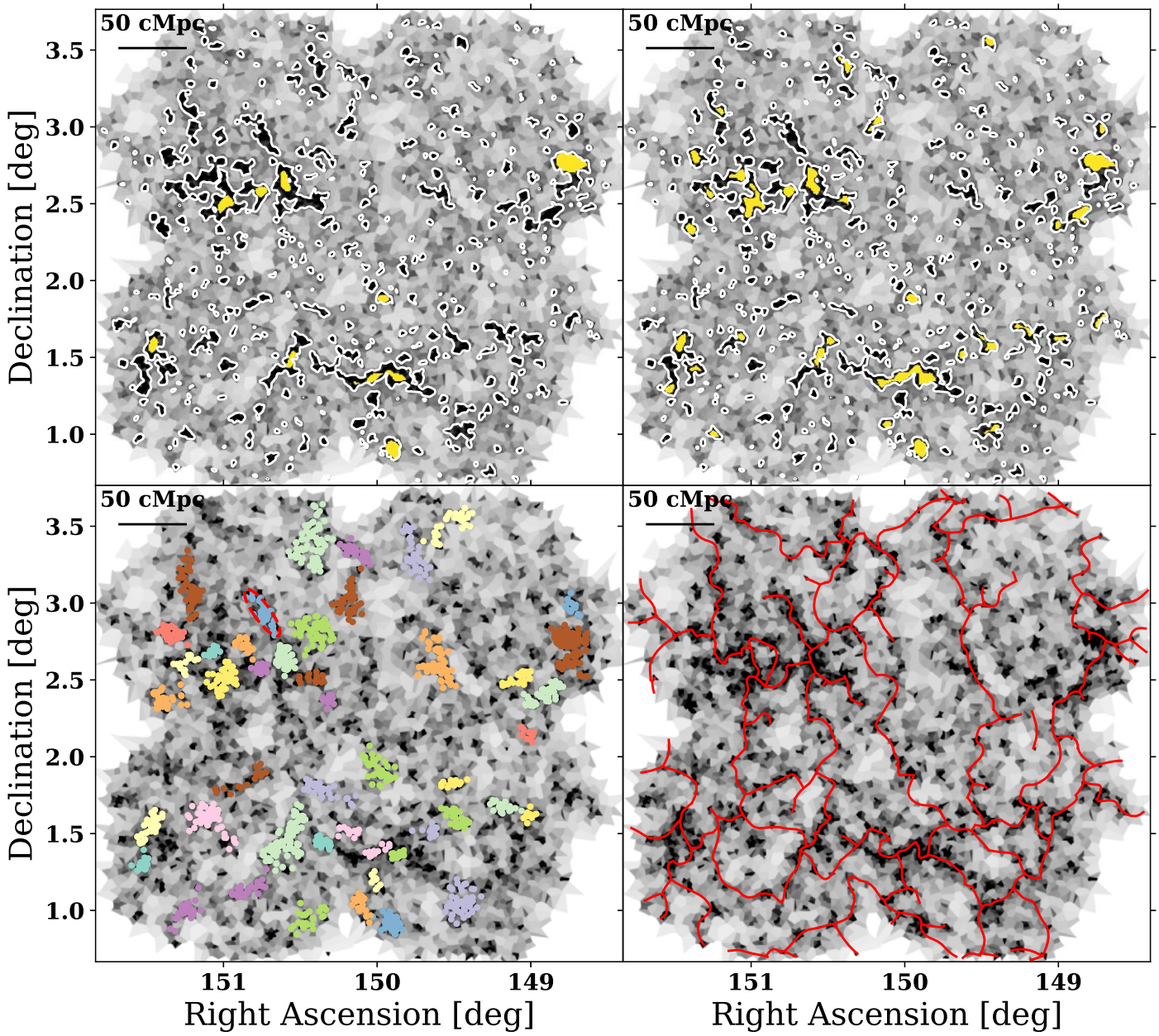

Figure 5. Protoclusters selected from the GS map (top left), VT map (top right), with HDBSCAN (bottom left) and filaments identified with DisPerSE with a persistence threshold of 2σ (bottom right) overlaid on the VT map. The red dashed region in the bottom left panel highlights a structure that is missed by the VT and GS maps but is detected with HDBSCAN. The filaments correspond to regions of moderate ($\sim 2\sigma$–3σ) overdensity, and connect protoclusters together.

reveals a complex web extending over the entire field. Several previously noted linear overdensities (e.g., within complexes A and C) are part of this network. Moreover, the filaments appear to converge at the locations of the extended complexes, interconnecting the individual protoclusters within. Indeed, every protocluster is connected to at least one filament, with several positioned at the convergence of multiple filaments. Our findings are fully consistent with the hierarchical picture wherein protoclusters occupy the nodes of the cosmic web.

Given that the line-of-sight window function set by the *N*501 filter is substantially larger than the dimensions of individual protoclusters and cosmic filaments, our 2D-based detection algorithm is expected to include false positives arising from random noise fluctuations. In Section 5.3, we show that, while false detection—in particular, of filaments—does indeed occur, all filaments around overdense protoclusters are robustly identified. Protocluster detection is even more secure.

5. Cosmic Structures in Simulations: Building Expectations with TNG

Cosmological hydrodynamical simulations, such as TNG300, offer invaluable insights not readily accessible through direct observations. By facilitating connections of observable traits of protoclusters and cosmic filaments with more fundamental attributes and allowing us to track their evolution across cosmic time, simulations provide a crucial context in which we understand our data. In this section, we compare the ODIN protocluster and filament samples with those derived from the TNG300 simulation to gain an

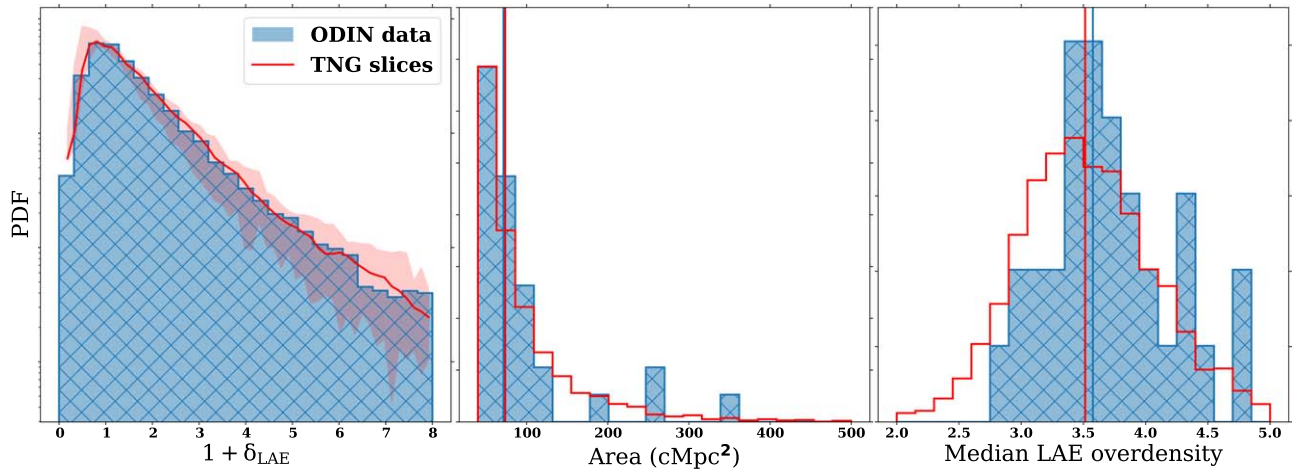

Figure 6. Left: LAE surface density distribution of the TNG300 slices, in comparison to that of the observed LAEs. The red line shows the median distribution across all 90 slices while the shaded region shows the spread. The observed and simulated LAE distributions are very similar, particularly at the high end, which represents massive structures. Middle and right: area and median LAE overdensity ($\delta_{\rm LAE}$) for the protocluster candidates detected in the 90 TNG300 slices compared to those of the observationally detected protocluster candidates. Blue/red vertical lines indicate the median value for the ODIN data/TNG300. The properties of the structures are similar in the observations and simulation.

understanding of the physical properties of the structures detected in Section 4.

5.1. Comparison of Observations and Simulations

In the left panel of Figure 6, we compare the LAE surface density distribution of the real data with the TNG300 slices. As detailed in Section 2.2, the latter are constructed from TNG300 and are designed to match the ODIN filter transmission as well as the LAE surface density and clustering bias. Multiple realizations are averaged over to result in the surface density distribution shown for the simulation. The high-end tail of the LAE surface density distribution, which represents the field's highest-density regions corresponding to protoclusters and filaments, is nearly perfectly reproduced.

In Figure 7, we show the VT map of both data and one $z = 3$ TNG300 slice side by side. The displayed slice is chosen at random and is centered on a massive protocluster with a descendant mass of $5.5 \times 10^{14}\,M_{\odot}$ (group ID = 13 at $z = 0$: cyan cross). It also contains five additional structures (yellow crosses) that will evolve into a cluster with mass greater than 2×10^{14} by $z = 0$, as well as several objects that will evolve into less massive clusters (green crosses). Indeed, the majority of the largest overdensities seen within the slice are associated with protoclusters. The two maps are remarkably similar in that both display structures of similar sizes and irregular morphologies, arranged into extended complexes. We highlight three such similar regions (complexes A1, B1, and C1 in the real data, and A2, B2, and C2 in the simulations, shown by dashed-line boxes). These mock structures are also connected by cosmic web filaments.

The distributions of the total transverse area and median LAE surface density of protoclusters detected in real data and simulation span a similar range, demonstrating good agreement given the constraints of modeling. This is illustrated in Figure 6. The median number of protoclusters (averaged over 90 TNG slices) is 37 compared to 33 detected in our data. They correspond to the protocluster surface density of $(4.1 \pm 0.6) \times 10^{-4}$ and $(3.3 \pm 0.6) \times 10^{-4}\,\mathrm{cMpc}^{-2}$, respectively. The uncertainties represent the shot noise only and thus should be considered as a lower limit.

One caveat is that the possible effects of radiative transfer on the observability of LAEs are ignored. In practice, the presence of dense gas and dust, particularly in the cores of protoclusters, may hide the LAEs in these regions (e.g., R. Shimakawa et al. 2017; R. Momose et al. 2021; Y. Huang et al. 2022). This might lead us to either underestimate the LAE surface density in the cores of our observational protocluster candidates or incorrectly pinpoint their centers. Since the median densities of the observed and simulated LAEs are similar, we expect that any underestimation of LAE surface density is relatively small. However, we cannot rule out the possibility of mislocating the centers, and there is some evidence suggesting discrepancies between the locations of overdense regions identified by LAEs and those identified by Lyman Break galaxies (e.g., K. Shi et al. 2019). Nevertheless, the excellent agreement between our data and simulations suggests that the massive cosmic structures identified with current and future ODIN data are robust.

5.2. Calibration of Descendant Mass Estimate

Estimations of the descendant mass allow us to establish direct connections between high-redshift protoclusters and present-day galaxy clusters, thereby tracing the evolution of massive cosmic structures and their galaxy populations across cosmic time. Following the methodology outlined by C. C. Steidel et al. (1998), we compute the descendant mass of ODIN protoclusters by assuming that the mass contained within the overdensity region will collapse into a cluster-sized halo by $z = 0$, with a total mass $M_{z=0}$ given by

$$M_{z=0} = \rho_m V_{\rm PC} = \rho_{0,z}(1 + \delta_m) V_{\rm PC} \tag{4}$$

where ρ_m and δ_m are the density and overdensity of matter, respectively. $V_{\rm PC}$ is the protocluster volume, and $\rho_{0,z}$ is the mean density of the Universe at redshift z. If the galaxy bias, b_g, is known, Equation (4) becomes

$$M_{z=0} = \left(1 + \frac{\delta_g}{b_g}\right) \rho_{0,z} V_{\rm PC} \tag{5}$$

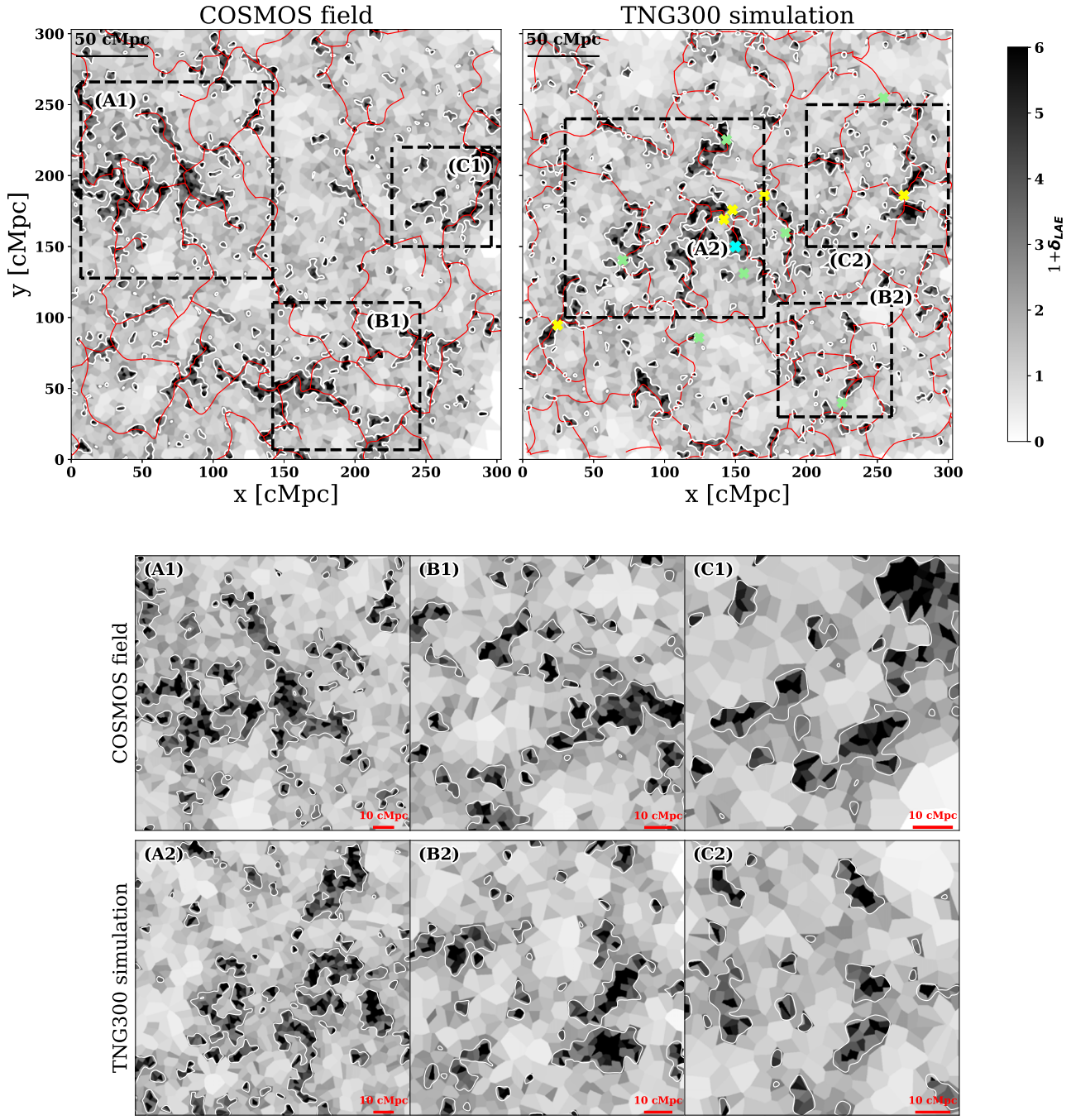

Figure 7. VT-based surface density map of our observed LAEs vs. that of a 80 cMpc thick slice of the TNG300 simulation box at $z = 3$. The positions of individual LAEs/mock LAEs are shown in Figure 1. Red lines in the top panels represent filaments selected with DisPerSE with a persistence threshold of 2σ. The cyan cross in the top right panel indicates the position of the massive protocluster on which the slice is centered (group ID = 13 at $z = 0$), while yellow crosses show the other massive protoclusters in the same volume (group IDs = 1, 2, 7, 11, 23 at $z = 0$). Green crosses show the progenitors of less massive clusters within the slice. Some overdensity complexes (A1, B1, C1; A2, B2, C2) are highlighted and zoomed in on the bottom panels. The two maps show visible similarities.

where δ_g is the galaxy overdensity. We assume $b_g = 1.8$ (E. Gawiser et al. 2007; M. White et al. 2024) and fix δ_g to the median LAE overdensity within the protocluster region. Since we do not have any redshift information for the majority of our LAEs, we assume that the size of a protocluster in the line-of-sight direction is similar to that in the transverse direction and approximate the volume of a protocluster from its area ($A_{\rm PC}$) as $V_{\rm PC} = A_{\rm PC}^{1.5}$: i.e., the shape of a protocluster is approximated as a cube. $M_{z=0}$ is then estimated as

$$M_{z=0} = \left(1 + \frac{\delta_g}{b_g}\right)\rho_{0,z}\, A_{\rm PC}^{1.5}. \quad (6)$$

The relation $V_{\rm PC} = A_{\rm PC}^{1.5}$ is a good estimate for an isotropic structure. As we have already seen, many of our protoclusters are strikingly nonspherical in shape, and thus, the above relation may be a poor approximation. If a protocluster is

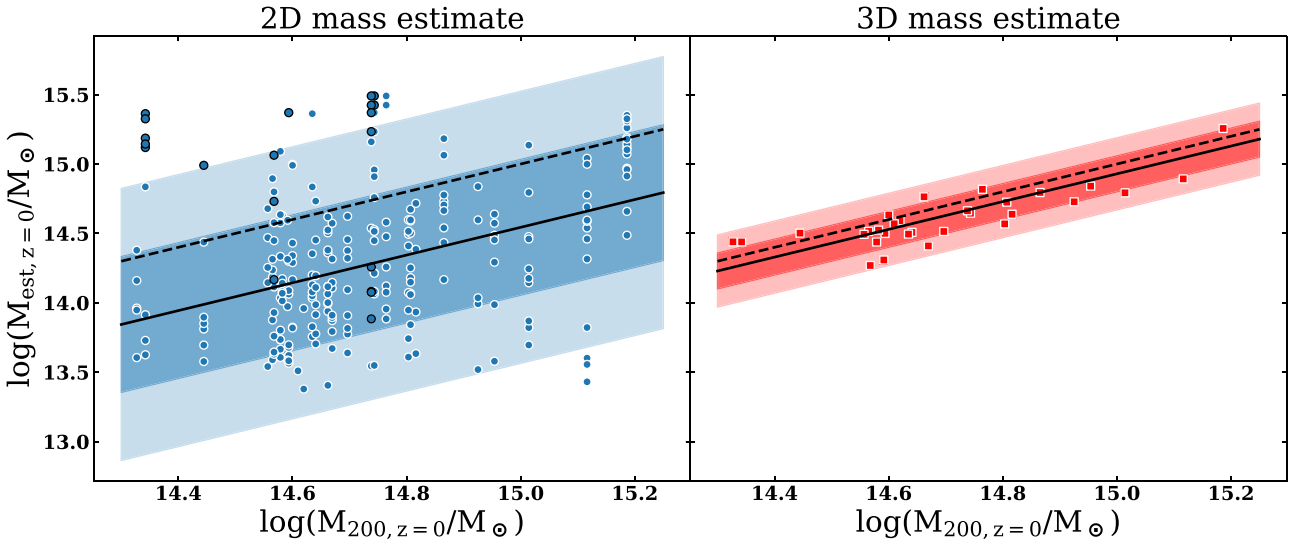

Figure 8. $M_{z=0}$ values estimated from Equation (6) compared to the true values for the progenitors of the 30 most massive clusters in the TNG300 box. The left panel (blue points) indicates the masses estimated from the 2D projection of the LAE overdensity corresponding to the protocluster, with the LAEs matched in number density to observations. The dashed line indicates a one-to-one relation between estimated and true descendant masses, while the solid line indicates a relation of $M_{\mathrm{est},z=0} = 0.35 M_{200,z=0}$. Highlighted points show cases where the protocluster was merged with another, more massive object by the detection. The right panel (red points) indicates the masses estimated from the 3D positions of the cluster members (see Section 5.2 for more details). Shaded regions indicate the range of the 1σ and 2σ scatter of the estimated masses. While the 2D mass estimate is subject to considerable scatter, primarily arising from the loss of information in the z-direction, it correlates well with the true descendant mass.

extended in the transverse dimension, it would be observed as a more modest overdensity with a larger angular extent. Conversely, a structure stretched along the line-of-sight direction is expected to be more compact with a larger overdensity. Using TNG-detected structures, we quantify the detection rate and the uncertainty in mass estimates due to the lack of information in the third dimension.

Figure 6 shows that there is a good agreement between TNG-selected and observational protoclusters for both angular sizes (middle) and median LAE overdensity (right). We also assess the likelihood of the 30 most massive cluster progenitors to be selected as a protocluster candidate when viewed along the X-, Y-, and Z-directions. To this end, we only consider the structures for which (i) the measured (angular) position center lies within 10 cMpc of the true center; and (ii) the line-of-sight position center is within 25 cMpc of the center of a given TNG slice. For reference, the FWHM of all slices is 60 cMpc, and a typical size of a protocluster is $\approx$10 cMpc.

Based on this definition, the median recovery rate of the protoclusters is 60%. The progenitor of the most massive cluster is always detected while the progenitors of the next four most massive clusters are detected >90% of the time. As expected, the likelihood of finding smaller structures depends on the sightlines being favorable or adverse to robust detection. If we lower the threshold for our VT-based protocluster selection from 4.5σ to 3.5σ, the recovery rate of the protoclusters increases to 80%. The higher success rate comes at the price of much greater contamination of $\sim$45%, compared to $\approx$20% for our fiducial setup.

In the left panel of Figure 8, we compare the *estimated* and true descendant mass, $M_{\mathrm{est},z=0}$ and $M_{200,z=0}$ for the recovered clusters in the 90 slices, where M_{200} refers to the mass contained within a spherical region whose mean density is 200 times the critical density of the Universe. The figure illustrates that $M_{\mathrm{est},z=0}$ tends to underestimate the true mass about 80% of the time. The median $M_{\mathrm{est},z=0}/M_{\mathrm{true},z=0}$ ratio is $0.35^{+0.67}_{-0.23}$, as indicated by the solid black line. This implies that the cosmic volume that ends up in a galaxy cluster by $z = 0$ is much greater than what we identify in the data as significant galaxy overdensities likely associated with a protocluster.

Our findings are consistent with the results from Y.-K. Chiang et al. (2013), who, based on the semianalytical models implemented in the Millennium simulations, found that about 40% of the total mass at $z = 0$, M_{today} is enclosed within a 6 (8) cMpc radius sphere for Virgo- and Coma-sized protoclusters with $M_{\mathrm{today}} = (3\text{–}10) \times 10^{14} M_{\odot}$ and $>10^{15} M_{\odot}$, respectively. In comparison, the median effective radius of our simulated protoclusters (computed as $(A_{\mathrm{PC}}/\pi)^{0.5}$) is $\sim$5 cMpc. To account for this effect, we correct the estimated masses by a factor of 0.35.

Figure 8 shows that the 2D-based mass estimates yield a considerable scatter of 0.49 dex. The most extreme outliers arise from cases where two or more protoclusters are merged together during detection, particularly for the protoclusters with $M_{200,z=0} \leqslant 10^{14.5} M_{\odot}$; these points are highlighted with a black outline in Figure 8. Even excluding these cases, the scatter remains high at $\sim$0.44 dex. We ascertain that the scatter is a result of the loss of information in the z-direction as follows. We perform the 3D VT of the 3D volume centered on the 30 massive clusters, this time using *all* the galaxies with $M_{\mathrm{star}} > 10^{7} M_{\odot}$. Using the subhalo merger trees of the $z = 0$ cluster members, we also identify the "member galaxies" in the $z = 3$ snapshot. The LAE overdensity and the protocluster region are computed similarly to the 2D case, but Voronoi cells are now 3D polyhedra instead of 2D polygons, while galaxy overdensity is measured as the median LAE overdensity of the Voronoi cells containing member galaxies. V_{PC} is obtained by summing over these Voronoi cells.

The resultant mass estimate is shown in the right panel of Figure 8. As expected, the scatter of the 3D $z = 0$ mass estimates is considerably smaller, at $\sim$0.15 dex, than in the 2D case, once we have eliminated all uncertainties regarding the cluster membership and the information on the third dimension. The estimated descendant mass is also significantly less

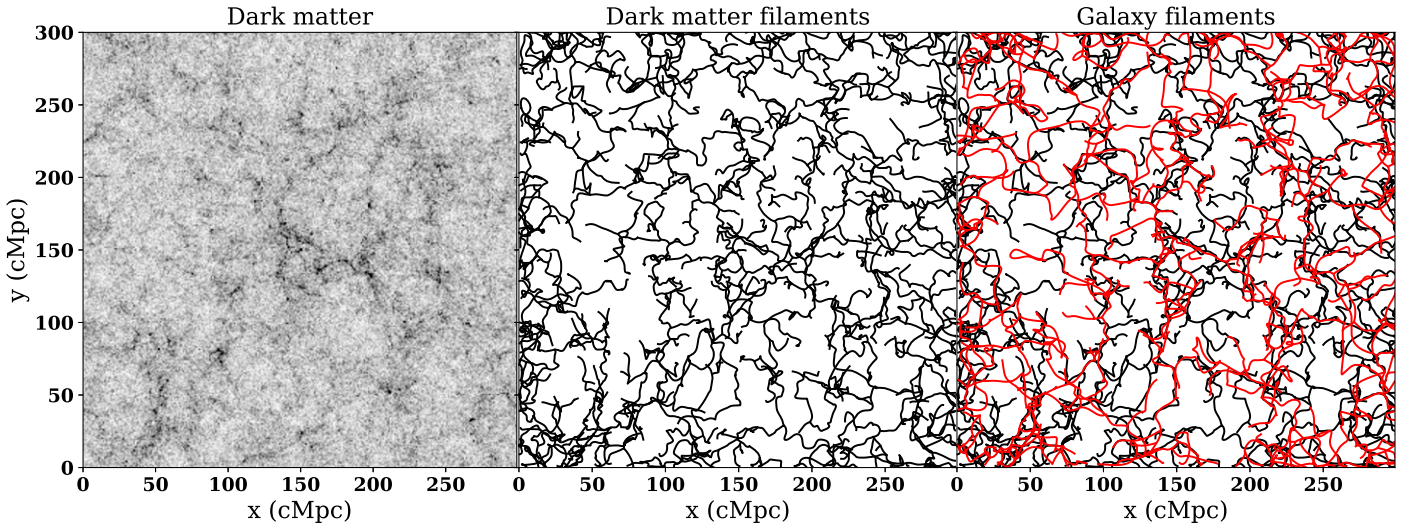

Figure 9. Left: distribution of dark matter in a randomly chosen TNG300 slice. Middle: the filament network identified with DisPerSE, using 0.1% of the dark matter particles and with a persistence threshold of 6σ. Right: the filament network identified with DisPerSE using galaxies (red), again with a persistence threshold of 6σ, overlaid on that traced by dark matter. Although the former is less detailed than the latter, there is excellent correspondence between the two.

underestimated than in the 2D case, being ∼85% of the true value. It is noteworthy that the *intrinsic* scatter of 0.15 dex is comparable to 0.2 dex estimated by Y.-K. Chiang et al. (2013) in converting galaxy overdensity measured within a $(15\ \mathrm{cMpc})^3$ cubic window into descendant mass.

When we compare the "true" volume found with the 3D VT to that derived from the projected protocluster area, we find that the scatter of the latter (≈0.52 dex) is sufficient to account for the entirety of the uncertainty in the mass estimate. This suggests that the uncertainty in $M_{z=0}$ is dominated by the volume estimation and that the isotropic assumption is less than ideal for protoclusters at high redshift. This result is not surprising, but it does quantitatively demonstrate the need for large-scale spectroscopic follow-up of protocluster candidates. However, we also emphasize that the 2D mass estimation is correct in an *average* sense, and statistical analyses for a large sample of protocluster candidates that make use of this mass estimate will be robust.

5.3. The Fidelity of Filament Recovery

At high redshift, active star formation and galaxy growth taking place in protoclusters are likely supported by cool gas transported along filaments of the cosmic web (e.g., A. Dekel et al. 2009; E. Daddi et al. 2021). The wide area coverage of ODIN presents a unique opportunity to study protoclusters in concert with the surrounding cosmic web, provided that we can accurately recover the filaments. As they are even more extended structures than protoclusters, the effect of projection along the line of sight on filament recovery must be considered separately. In this section, we use the TNG300 slices to examine how well the observed 2D filament network corresponds to the true underlying 3D network. We focus specifically on filaments in the vicinity of protoclusters, to understand how likely it is for a filament that appears near a protocluster in projection to be physically connected to the protocluster.

We make use of the 90 slices of the TNG300 full box at $z = 3$ described in Section 2.2. For each slice, we use DisPerSE to extract filaments using the positions of the mock LAEs matched to ODIN, projected along the slice axis. This is henceforth referred to as the "2D skeleton" or "2D filaments." To create a comparison set of filaments that represents the "true" cosmic web, we also run DisPerSE on the 3D positions of the slice galaxies with $10^7\,M_\odot < M_* < 10^{12}\,M_\odot$, henceforth referred to as the "3D skeleton" or "3D filaments." We use a persistence threshold of 2σ to extract the 2D skeleton, identical to that used on the observed LAEs. For the 3D skeleton, we use multiple persistence thresholds of 5σ, 6σ, and 7σ to compare the 2D skeleton against 3D features of varying significance.

How well does the 3D skeleton traced by galaxies correspond to the underlying distribution of dark matter? S. H. Im et al. (2024) quantitatively analyze this question for the Horizon Run 5 simulation. They find that, while the filaments outlined by all the simulated galaxies are less detailed than the filaments traced by dark matter, the former nevertheless aligns closely with the latter. Although we do not attempt to replicate the full analysis of S. H. Im et al. (2024), we find similar behavior in TNG300, as demonstrated in Figure 9 where we show the filaments traced by dark matter and the 3D skeleton traced by galaxies in a randomly chosen slice.

In Figure 10, we show the dark matter distribution in one of the TNG300 slices overlaid with the 2D and 3D skeletons. The similarity between the DisPerSE filaments (both 2D and 3D) and the underlying dark matter distribution is clear, as is the fact that the 2D and 3D skeletons mirror each other closely. However, the 2D skeleton does not perfectly reproduce the true 3D cosmic web. To quantify the extent to which the cosmic web can be recovered by our observational data, we compare the 2D filaments to the 3D skeleton. We do this following a similar procedure to F. Sarron et al. (2019; see also C. Laigle et al. 2018; U. Kuchner et al. 2022), by matching the individual segments out of which DisPerSE constructs the 2D and 3D skeletons (see Section 4.3) to each other. We measure the (projected) separation between the midpoints of each segment of the 2D skeleton and those of the 3D skeleton. We denote the minimum distance *from* a segment of the 2D filament network *to* a segment of the 3D network as $d_{2d\rightarrow 3d}$, and the inverse as $d_{3d\rightarrow 2d}$. This mapping is illustrated in the left panel of Figure 11. The former measurement enables us to determine which of the 2D filaments represent true features of the 3D

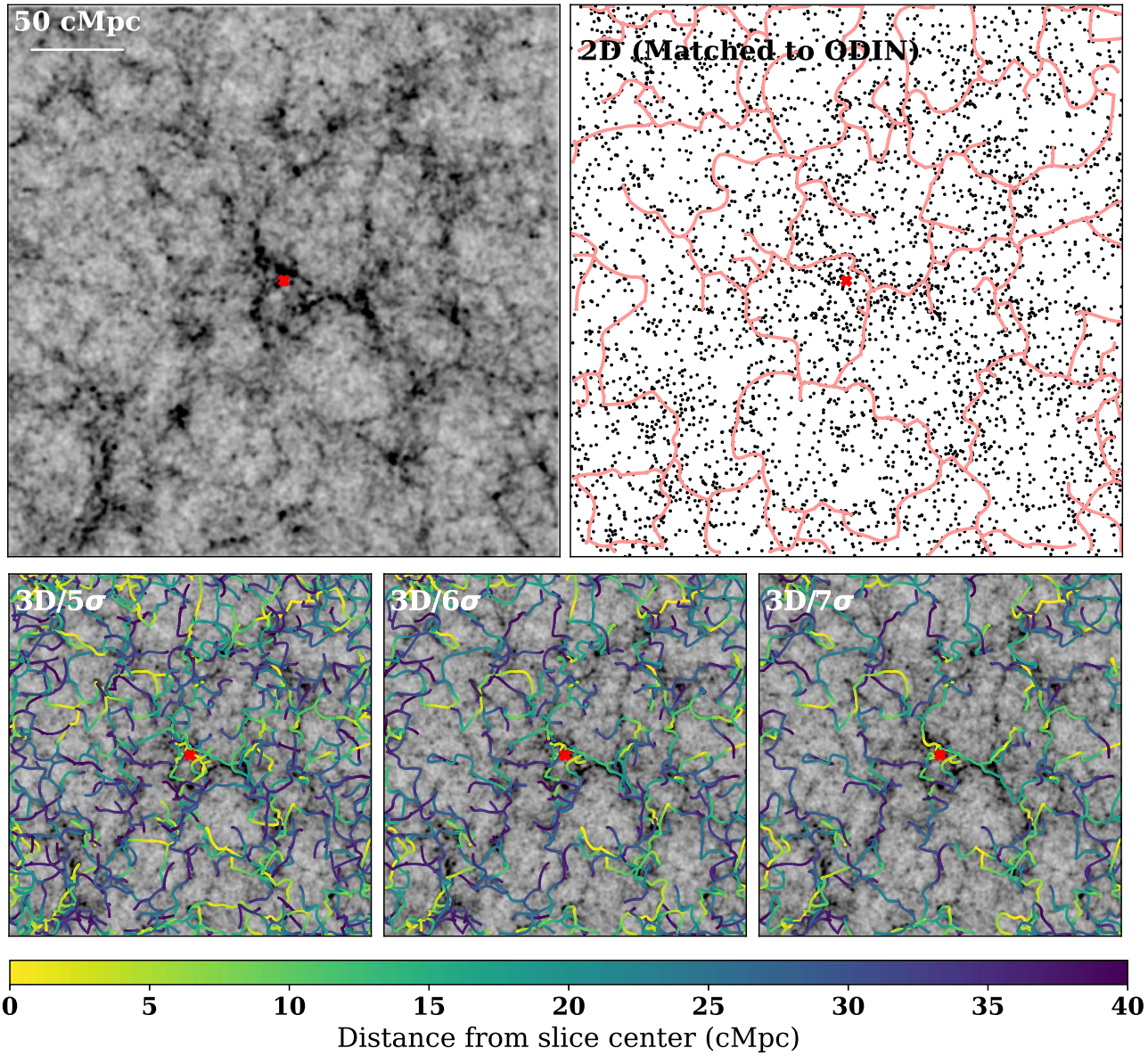

Figure 10. Top left: the dark matter distribution for a randomly selected slice, centered on a protocluster with descendant mass log $(M_{200}/M_{\odot}) = 15.18$ (red cross; group ID = 0 within the $z = 0$ TNG300 simulation box). Top right: the filaments (pink lines) extracted by DisPerSE for this slice are based on the mock LAE positions, which are shown by black dots. This network is comparable to the cosmic web selected with ODIN data. Bottom: the filaments extracted by DisPerSE for the same slice using the 3D positions of the slice galaxies with $10^7\,M_{\odot} < M_{*} < 10^{12}\,M_{\odot}$, with varying persistence thresholds. The filaments are color-coded by their separation from the center of the slice along the line-of-sight direction. At large distances from the slice center, the filaments are truncated by the slice boundaries. The number of such truncated filaments reduces at higher persistence thresholds.

cosmic web, while the latter allows us to determine which features of the underlying filament network are successfully recovered by observations.

In the top right panel of Figure 11, we show histograms of $d_{2d\rightarrow 3d}$ for the three persistence thresholds. Hydrodynamical simulations have found that the typical radius of a filament at $z \sim 3$ is 2–3 pMpc (8–12 cMpc; W. Zhu et al. 2021; see also S. H. Im et al. 2024). The median values of $d_{2d\rightarrow 3d}$ are well below this range, being 2.0, 2.7, and 4 cMpc respectively for the 3D skeletons found with persistence 5σ, 6σ, and 7σ. The values are less than 5 cMpc (1.25 pMpc) for the majority of the 2D segments. We refer to the segments with $d_{2d\rightarrow 3d} < 5$ cMpc as "matched 2D segments," meaning that they have a "match" in the 3D skeleton. The fraction of matched segments decreases with increasing persistence threshold, from ∼80% for a persistence threshold of 5σ to ∼60% when the persistence threshold is 7σ. This is reasonable; as the persistence threshold is raised, the extracted network is increasingly restricted to only the most significant filaments, and finer features are lost, leading to a greater number of 2D segments going unmatched. Irrespective of the adopted persistence value, 55% of the 3D segments have 2D matches within $d_{3d\rightarrow 2d} < 5$ cMpc. We conclude that the 2D skeleton serves as a good representation of the underlying cosmic web even though it faithfully recovers the individual features ≈55% of the time.

Next, we examine whether the fraction of matched 2D segments depends on the persistence threshold used to extract the 2D skeleton. If the matched fraction increases substantially with increasing persistence threshold, it would indicate that our chosen persistence threshold is too generous. Reassuringly, we

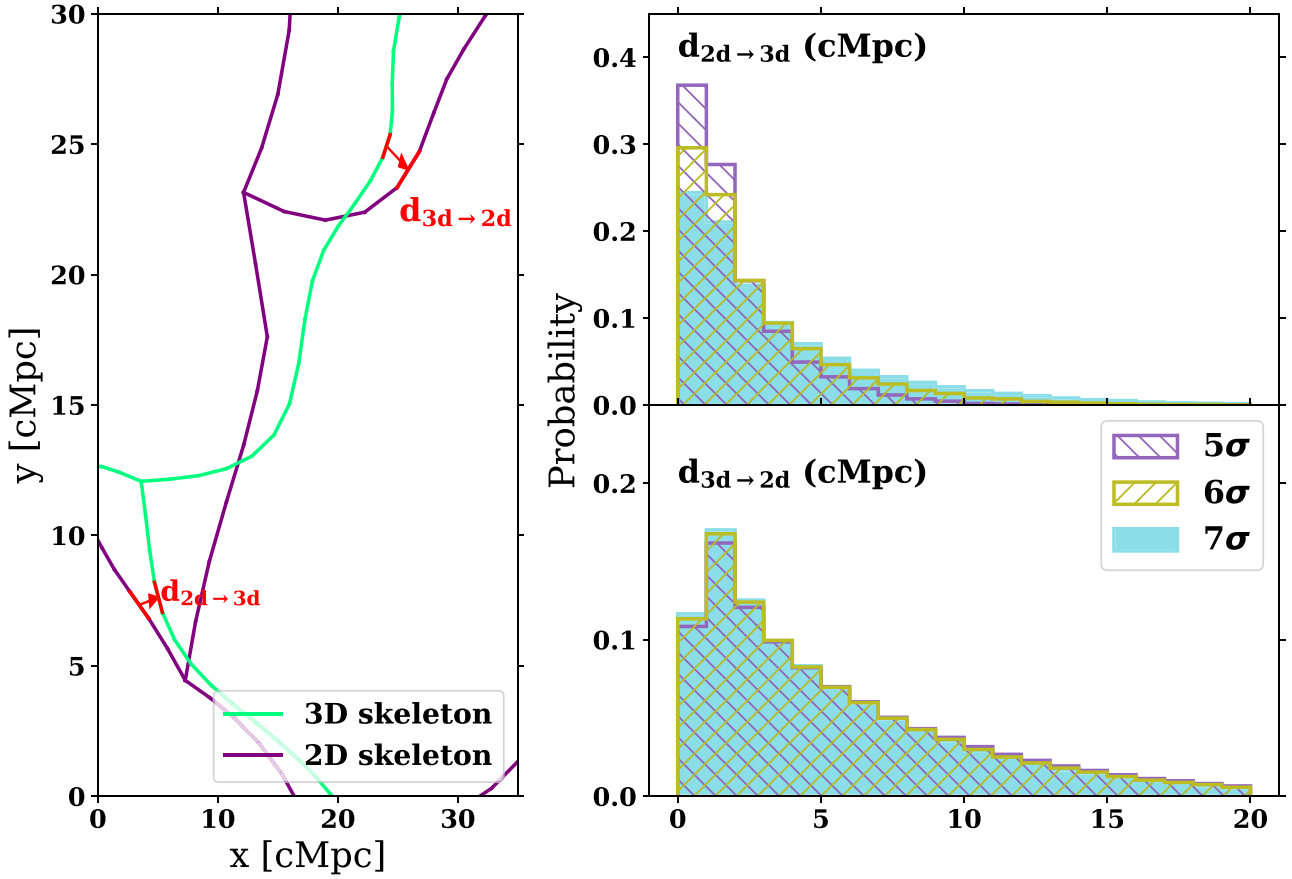

Figure 11. Left: an example comparison of the 2D and 3D skeletons, illustrating how $d_{2d\rightarrow 3d}$ and $d_{3d\rightarrow 2d}$ are measured. Right: comparison of the 2D and 3D skeleton using the separation of the segments, $d_{2d\rightarrow 3d}$ (top) and $d_{3d\rightarrow 2d}$ (bottom). The separations are calculated across all 90 slices of 80 cMpc thickness from the $z = 3$ TNG300 simulation. In the majority of cases, both $d_{2d\rightarrow 3d}$ and $d_{3d\rightarrow 2d}$ are less than 5 cMpc, regardless of the specific persistence threshold.

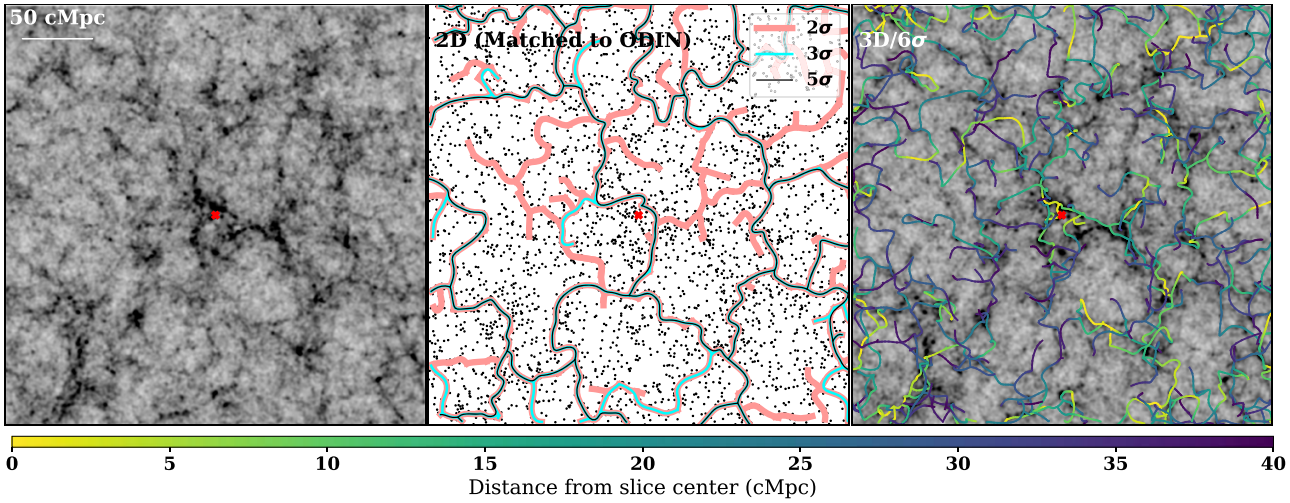

Figure 12. Similar to Figure 10, but showing the 2D skeletons extracted with multiple persistence thresholds.

find that the matched fraction only rises marginally with the persistence threshold, increasing by $\lesssim$5% from a persistence threshold of 2σ to 5σ. This is visualized in Figure 12, where we show the 2D skeletons extracted with persistence thresholds of 2σ, 3σ, and 5σ in comparison to the dark matter distribution and the 3D skeletons. The majority of the 2D filaments extracted with a lower persistence threshold, which do not appear with a higher one, can be visually matched to the true features of the 3D skeleton.

Intriguingly, the fraction of matched 2D segments increases with increasing surface density for all cases, as shown in the top panel of Figure 13, suggesting that the higher the observed LAE surface density is, the more likely the detected filament is to be an accurate representation of the 3D cosmic web. We also examine the converse, i.e., the probability that a filament within an overdense region is recovered by observations. On considering the 3D skeleton within 10 cMpc of a massive protocluster, we find that the fraction of matched 3D segments

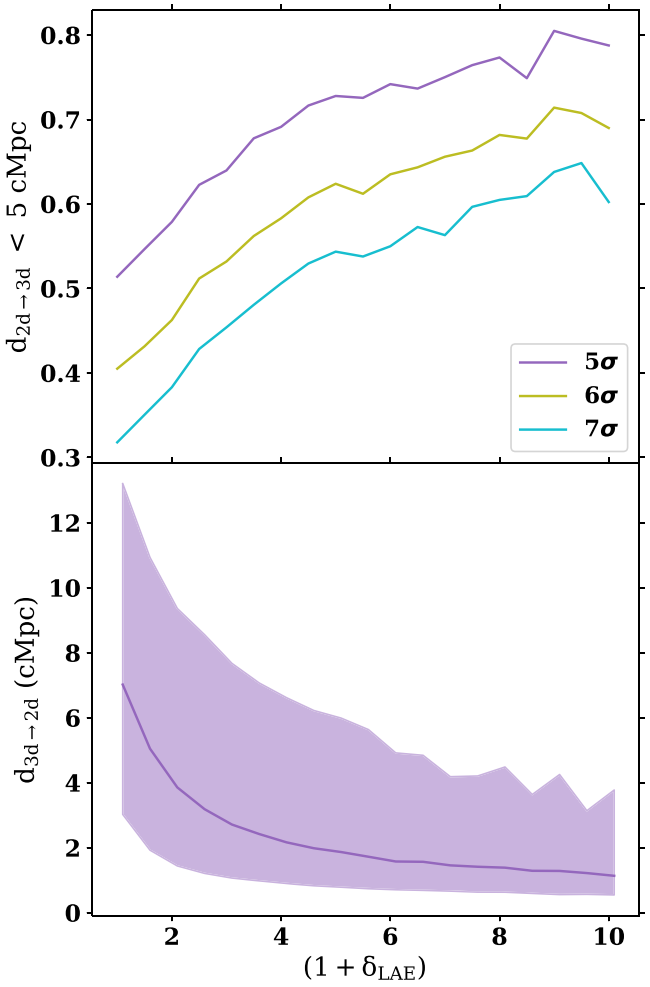

Figure 13. Top: the fraction of 2D filament segments with a match with the 3D skeleton (detected with various persistence thresholds), as a function of the LAE surface density at the position of the filament segment. Regardless of the persistence threshold used to select the 3D skeleton, the fraction of matched 2D segments increases with increasing surface density. Bottom: $d_{3d\to 2d}$ as a function of the (2D) LAE surface density for the 5σ threshold 3D skeleton; the other two skeletons give virtually identical results. The solid line indicates the median value while the shaded region indicates the spread as measured from the 16–84 percentile scatter.

is considerably higher than in average regions, increasing from a median of ∼55% to a median of ∼86% in all three persistence cases. The matched fraction is even higher if we restrict the region under consideration to within 5 cMpc of the protoclusters, being 100% in the majority of slices.

In the bottom panel of Figure 13, we show $d_{3d\to 2d}$ as a function of the (2D) LAE surface density in the slices. To avoid confusion, we only show the result for the 5σ threshold skeleton, as the remaining two persistence thresholds give nearly identical results. With increasing LAE surface density, $d_{3d\to 2d}$ decreases rapidly; i.e., in dense protocluster regions, not only is the recovery rate of the 3D skeleton considerably higher but also the 2D skeleton resembles the 3D skeleton more closely. This is illustrated in Figure 14 where we show the 2D and 3D skeletons for the same slice as in Figure 10 but zoom in on the massive protocluster. These results show that the ODIN data will be capable of studying protoclusters within the context of the surrounding cosmic web.

6. Discussion

6.1. Comparison of Structure Detection Methods

In Section 4, we based our protocluster selection on the GS and VT maps and the HDBSCAN algorithm. These methods resulted in 9, 33, and 47 structures, respectively. We now wish to evaluate the similarities and differences between these methods.

The false positive rate in the GS method is very low for all sets of detection parameters and is ∼0 for our fiducial criteria. The nine GS-based protoclusters may thus be considered the most secure candidates. All GS structures are present in the VT sample with similar morphologies. In all cases, the separation between the centers determined by the GS and VT method is within 2.5 cMpc. For all but one that is present in both GS and VT samples, the measured area is slightly greater (by a factor of ∼1.1–1.8) in the latter. This is because irregular overdensities are generally detected at higher significance in the VT map.

All structures in the GS sample are also present in the HDBSCAN sample. One is split into two structures by HDBSCAN as shown in the middle column of Figure 4. For the remaining eight, the structures have similar morphologies in both the GS and HDBSCAN analyses; however, the measured area in HDBSCAN is larger than that estimated from the GS method by a factor of ∼2–5.

More interesting is the comparison between the structures detected from the VT map and using HDBSCAN. Figure 5 shows the locations of the VT (upper right) and HDBSCAN samples (lower left). Only two of the 33 protocluster candidates in the VT sample are not detected by HDBSCAN. As for the remaining 31, one is recovered as two separate objects and two are merged into a single structure. For all the protoclusters detected by both the VT and HDBSCAN methods, the latter yields larger measured areas than the former, typically by more than a factor of 2.

Why are the VT-based structures smaller and fewer in number compared to the HDBSCAN-selected ones? Visual inspection of Figure 5 suggests that many HDBSCAN-selected structures with no VT counterpart are more elongated and filamentary. In particular, the filamentary arm previously highlighted in complex A is identified by HDBSCAN (highlighted in red in the figure) but not from the VT map. If we relax the overdensity threshold for the VT detection from 4.5σ to 3.5σ, the number of HDBSCAN-selected protoclusters that overlap one or more VT-based ones increases from 30 to all 47. The areas of the two sets of protoclusters become more similar. However, the downside of lowering the detection threshold for the VT map is the increased contamination rate. We conclude that HDBSCAN fares better in recovering lower surface density features with a more elongated morphology, but with the downside that it is more difficult to ascribe a physical meaning to the detection parameters for this method (Kth nearest neighbor and minimum cluster size) than for those for the GS and VT methods (minimum area and density threshold).

6.2. Descendant Masses of the Protocluster Candidates

The number densities of the VT and HDBSCAN protocluster samples (which we expect to be more complete than the GS sample) are 5.5×10^{-6} cMpc^{-3} and 7.9×10^{-6} cMpc^{-3}, respectively. The number density of clusters at $z=0$ in TNG300 is 1.2×10^{-5} cMpc^{3}, implying that our protocluster

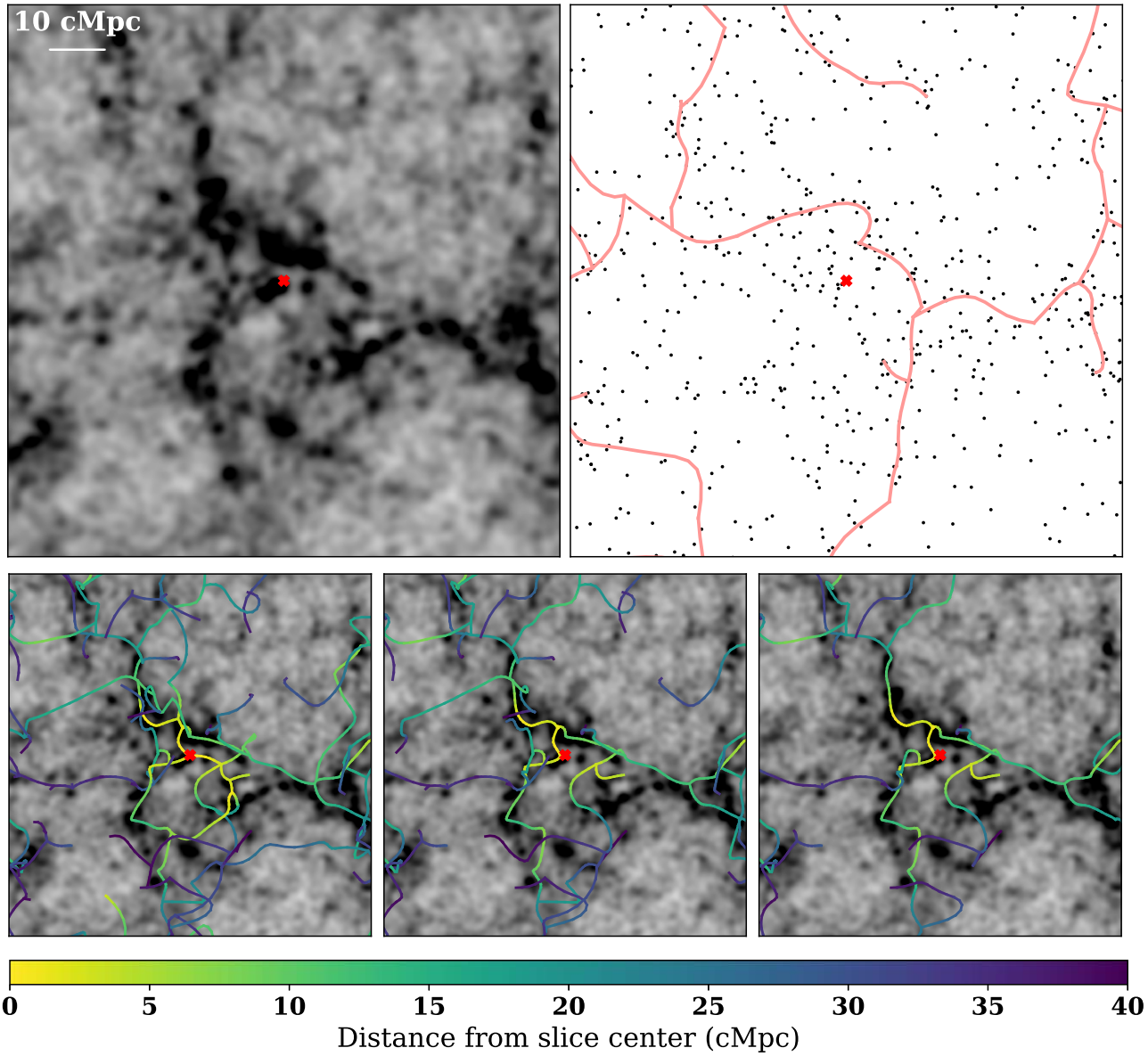

Figure 14. As Figure 10, but zoomed in on the region surrounding a massive protocluster.

sample has a completeness of ∼50%. Purely based on the comoving number density, i.e., by assuming that each of our protocluster candidates evolves into an independent galaxy cluster by $z = 0$, the descendant masses are estimated to be $\gtrsim 2 \times 10^{14}\,M_{\odot}$ and $1.6 \times 10^{14}\,M_{\odot}$, respectively. In practice, some of our structures will likely merge into a single more massive cluster. Thus, the masses quoted above represent a lower limit.

We now estimate the descendant masses of individual protocluster candidates. We showed in Section 5.2 that, for our choice of protocluster detection parameters, Equation (6) underestimates the descendant masses of the structures selected from the VT map by a factor of ∼3. A similar analysis for the GS and HDBSCAN methods (see Appendix B) shows that the masses of the former are underestimated by a factor of ∼2, and those of the latter are underestimated by a factor of ∼1.15. In Figure 15, we show histograms of the descendant masses after making the appropriate corrections. The median masses are $\log(M_{z=0}/M_{\odot}) = 14.35$, 14.75, and 14.52 for the VT, HDBSCAN, and GS protoclusters, respectively, suggesting that our protocluster candidates will evolve into moderately massive "Virgo-type" clusters (Y.-K. Chiang et al. 2013). The median values and the overall range spanned by the descendant masses are in reasonable agreement with the lower limits inferred from the number density. This again highlights the completeness of our protocluster sample.

6.3. Comparison with Other Works

Several studies have made use of LAEs to trace the LSS at $z \sim 2$–4 (e.g., T. Hayashino et al. 2004; T. Yamada et al. 2012; K.-S. Lee et al. 2014; T. Bădescu et al. 2017; S. Kikuta et al. 2019; Y. Huang et al. 2022). ODIN reaches a comparable Lyα line flux to these past surveys (~ 1–9×10^{-17} erg s^{-1} cm^{-2}), but over a considerably larger contiguous area—T. Yamada et al. (2012) and S. Kikuta et al. (2019) have the largest area coverage at 1.38 and 1.1 deg^2, respectively, while the remainder survey areas of ∼0.2–0.5 deg^2. ODIN is also unusual in that our protocluster candidates are discovered blindly, with no prior information available; T. Bădescu et al. (2017) and

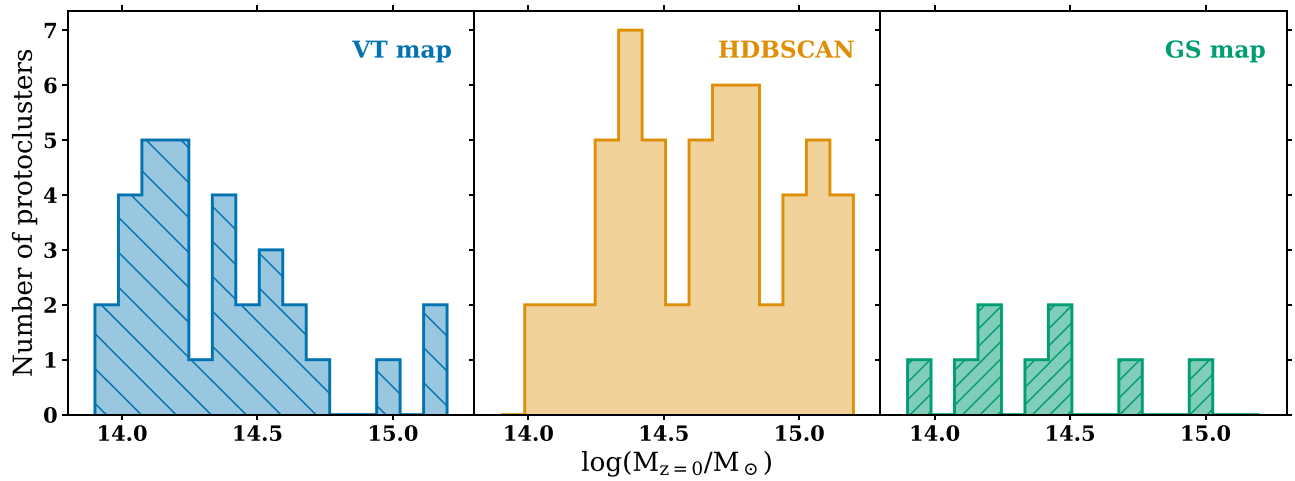

Figure 15. Descendant masses of the individual protoclusters for all three samples, after applying appropriate corrections to the mass calculated with Equation (6).

S. Kikuta et al. (2019) target the vicinity of a pair of Lyα nebulae and a hyperluminous QSO, respectively, while T. Hayashino et al. (2004), T. Yamada et al. (2012), K.-S. Lee et al. (2014), and Y. Huang et al. (2022) used narrowband imaging to follow up on previously identified galaxy overdensities.

It is worth noting that the structure studied by T. Yamada et al. (2012), located at $z = 3.09$ in the SSA22 field, is strikingly similar in its size and clumpy morphology to our complex A. The entirety of the 1.38 deg^2 region they survey is overdense in LAEs, of which a region of $\sim$60 cMpc^2 was previously studied by C. C. Steidel et al. (2000) and T. Hayashino et al. (2004) and comprises the core of the protocluster. The core region has been shown to host multiple extended and luminous Lyα nebulae (Y. Matsuda et al. 2004), a trait shared by complex A (see V. Ramakrishnan et al. 2023), suggesting that the two systems are similar.

The protocluster studies using different techniques include J. Toshikawa et al. (2016), who used *u*-dropout galaxies to identify five protoclusters at $z \sim 3$ over 4 deg^2 of the fields covered by the Canada–France–Hawaii Telescope Legacy Survey (M. Sawicki et al. 2019) and J. Toshikawa et al. (2018), who used *g*-dropout galaxies to find 216 protoclusters at $z \sim 3.8$ over the 200 deg^2 of the HSC-SSP Wide fields. These samples have number densities of $\sim 1 \times 10^{-7}$ and 1.5×10^{-6}, respectively. Y.-K. Chiang et al. (2014) use galaxies with photometric redshifts from the COSMOS/UltraVISTA catalog to find 26 protoclusters at $1.6 < z < 3.1$ over an area of $\sim$1.2 deg^2, corresponding to a number density of 1.2×10^{-6} cMpc^{-3}. These works identify overdensities of similar sizes (5–15 cMpc) and, in the case of J. Toshikawa et al. (2018), similar descendant mass ($\sim$1–2 $\times 10^{14}\, M_\odot$). However, the redshift uncertainty for the tracers used by these works is much larger than that of our LAEs. This may lead to density peaks being diluted by interlopers, resulting in lower number densities.

7. Conclusions

The ODIN survey is the largest-area deep field narrowband survey undertaken to date. By enabling the selection of LAEs, which are low-mass, star-forming galaxies and are well localized in redshift space over a wide contiguous area, ODIN makes it possible to comprehensively trace the LSS at high redshift. In this paper, we have used the early ODIN data taken in the COSMOS field with the *N*501 filter to compile a sample of protoclusters and cosmic filaments.

We create LAE surface density maps using two methods—by smoothing over the LAE positions with a fixed-size Gaussian kernel (GS map, Section 3.1) and by constructing the Voronoi diagram of the LAEs (VT map, Section 3.2). We select protocluster candidates from the surface density maps (Section 4.1) and by applying the density-based clustering algorithm HDBSCAN to the LAE positions (Section 4.2). We also select filaments of the cosmic web (Section 4.3). We assess the reliability of our structure detection by comparing our observations against the results obtained with a carefully created mock sample of LAEs from the IllustrisTNG300-1 hydrodynamical simulation (Section 5). Our main results and conclusions are as follows:

1. The LSS revealed by our surface density maps is distinctly clumpy and irregular, with overdense regions clustered together in extended complexes. The VT map identifies filaments as regions of moderate LAE overdensity ($(1 + \delta_{\rm LAE}) = 2$–3) connecting the regions of highest density ($(1 + \delta_{\rm LAE}) > 3$). These features are in line with the expectations of hierarchical structure formation.
2. The three methods of identifying protoclusters that we explore all have their strengths and weaknesses. The VT map identifies a greater number of objects but also suffers from a higher contamination rate than the GS map. The GS map recovers only the most significant overdensities but is almost free of interlopers. The HDBSCAN map recovers the maximum number of structures; however, the input parameters are less straightforward to interpret or assign physical meaning to than for the surface density maps.
3. The surface density maps of the $z = 3$ TNG300 simulation box display remarkably similar features to those seen in the observations. The number and size of the overdensities selected with our fiducial detection parameters are likewise similar in the two. As shown in Figure 7, many of the prominent overdensities in the TNG300 surface density maps correspond to the progenitors of $z = 0$ clusters.

4. The simulation shows that we can successfully recover ∼60% of the protoclusters with a descendant mass $\gtrsim 2 \times 10^{14} M_{\odot}$. We find that the descendant mass of the simulated protoclusters can be estimated purely based on their area and LAE overdensity measured in 2D. The lack of information in the redshift direction introduces a scatter of ∼0.4 dex on the measurement.
5. By comparing the filaments identified in 2D and 3D in TNG300, we show that, despite projection effects, the cosmic web recovered in 2D is a close representation of the true LSS in regions with high LAE surface density. In the vicinity of protoclusters, the 3D filament network is almost perfectly recovered.
6. On estimating the descendant masses of our observed protocluster samples, we find that they span the range log $(M_{z=0}/M_{\odot}) \sim 14.0$–15.0, with the median of the VT, GS and HDBSCAN samples being 14.35, 14.52, and 14.75 respectively. The majority of our protoclusters are thus likely to evolve into intermediate-mass "Virgo-type" clusters.

Our results establish the robustness of our protocluster and filament samples and demonstrate that LAEs are reliable and efficient tracers of LSS at high redshift. Upon completion, the ODIN survey will allow us to select secure samples of massive structures that are nearly 10 times larger than those presented here and span three cosmic epochs. ODIN will thus be well placed to grant insight into the formation and evolution of cosmic structures near Cosmic Noon.

Acknowledgments

The authors acknowledge financial support from the National Science Foundation under grant Nos. AST-2206705, AST-2408359, and AST-2206222, and from the Ross-Lynn Purdue Research Foundations. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under grant No. DGE-2233066 to NF. M.C.A. and A.K. acknowledge financial support from ANID/Fondo 2022 ALMA/31220021. M.C.A. acknowledges support from ANID BASAL project FB210003. The Institute for Gravitation and the Cosmos is supported by the Eberly College of Science and the Office of the Senior Vice President for Research at the Pennsylvania State University. H.S.H. acknowledges the support from the National Research Foundation of Korea grant funded by the Korean government (MSIT; No. 2021R1A2C1094577). J.L. is supported by the National Research Foundation of Korea (NRF-2021R1C1C2011626). S.L. is supported by the National Research Foundation of Korea grant funded by the Korean government (MSIT; Nos. 2020R1I1A1A01060310, 2020R1A2C3011091, 2021M3F7A1084525). Based on observations at Cerro Tololo Inter-American Observatory, NSF's NOIRLab (prop. ID 2020B-0201; PI: K.-S. Lee), which is managed by the Association of Universities for Research in Astronomy under a cooperative agreement with the National Science Foundation.

Facility: Blanco.

Software: SExtractor (E. Bertin & S. Arnouts 1996), SEP (K. Barbary 2016), hdbscan (L. McInnes et al. 2017), Astropy (Astropy Collaboration et al. 2013, 2018, 2022).

Appendix A
Parameter Selection for HDBSCAN

As stated in Section 4.2, HDBSCAN is a density-based clustering algorithm that separates points into "clusters" and "noise." HDBSCAN performs this classification by estimating the density at each data point from its Kth nearest neighbor distance, referred to as the "core distance." Points in dense (sparse) regions will have relatively small (large) core distances, and the core distance of a data point is thus inversely proportional to its density.

HDBSCAN is an extension of the DBSCAN algorithm. DBSCAN selects clusters by imposing a user-defined density threshold. This is implemented by placing a ceiling on the core distance/Kth nearest neighbor distance, that is, by requiring that there be K points within a distance ϵ or less of a given point. DBSCAN begins with a point that satisfies the requirement of having a core distance less than or equal to ϵ and classifies it and its neighbors within ϵ as a single cluster. If any of these neighbors also pass the density criterion, their neighbors within ϵ are assigned to the same cluster as well. This continues until there are no more points within the cluster that have a core distance less than ϵ, at which point the cluster is closed. DBSCAN then selects an unassigned point with K or more neighbors within ϵ and repeats the process. This continues until no points that pass the density threshold remain, at which juncture all remaining data points are classified as noise.

The approach of DBSCAN, of applying a constant density/core distance threshold, is restrictive. The optimal value of ϵ is difficult to identify and can vary from region to region within the data. HDBSCAN circumvents the difficulty of selecting a single density threshold by making use of *hierarchical clustering*; that is, it creates a hierarchy of clusters by applying successively higher-density thresholds and then selects the optimal set of clusters from this hierarchy. This makes it ideal for selecting clusters of varying density from a single data set, as in our case.

As it constructs a cluster hierarchy, HDBSCAN does not require the user to supply the parameter ϵ. Instead, the parameters that govern the cluster selection are K, as before (which determines how density is measured) and an additional parameter, the minimum cluster size. HDBSCAN uses the minimum cluster size to condense the cluster hierarchy—any cluster in the hierarchy that has fewer points than this size is discarded. This is shown in Figure 16 for a sample data set. From this condensed hierarchy, HDBSCAN selects the final set of clusters.

The HDBSCAN library provides two ways of selecting the final set of clusters. The first method, the "excess-of-mass" method, selects the most stable set of clusters, defined as those that persist for the greatest range of density thresholds. The second method, the "leaf" method, selects the smallest surviving clusters in the hierarchy. These two methods are illustrated in Figure 17 for the same sample data set.

Based on the above description, there are two main parameters that govern the clustering based on HDBSCAN, namely, K and the minimum cluster size. By determining which nearest-neighbor distance will be used for estimating the density, K controls how finely the density distribution is sampled, in essence acting as a kind of smoothing scale. Changing the minimum cluster size affects the condensed cluster hierarchy by changing the number of points required to be considered a cluster, that is, more points are discarded as

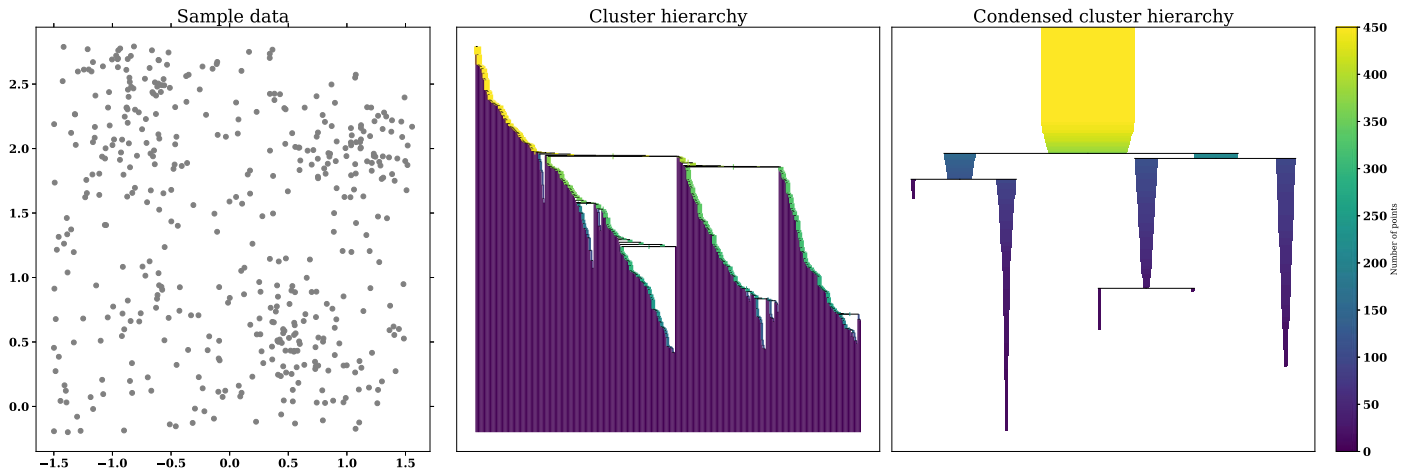

Figure 16. The cluster hierarchy (middle panel) and condensed cluster hierarchy (right panel) constructed by HDBSCAN for a sample data set (left panel). The sample data set contains three regions of densely clustered points along with random noise. The cluster hierarchy is constructed by applying successive density cuts. Each branch of the tree represents a cluster of points selected above some density cut. The condensed cluster hierarchy shows the cluster hierarchy after clusters smaller than the minimum cluster size are discarded.

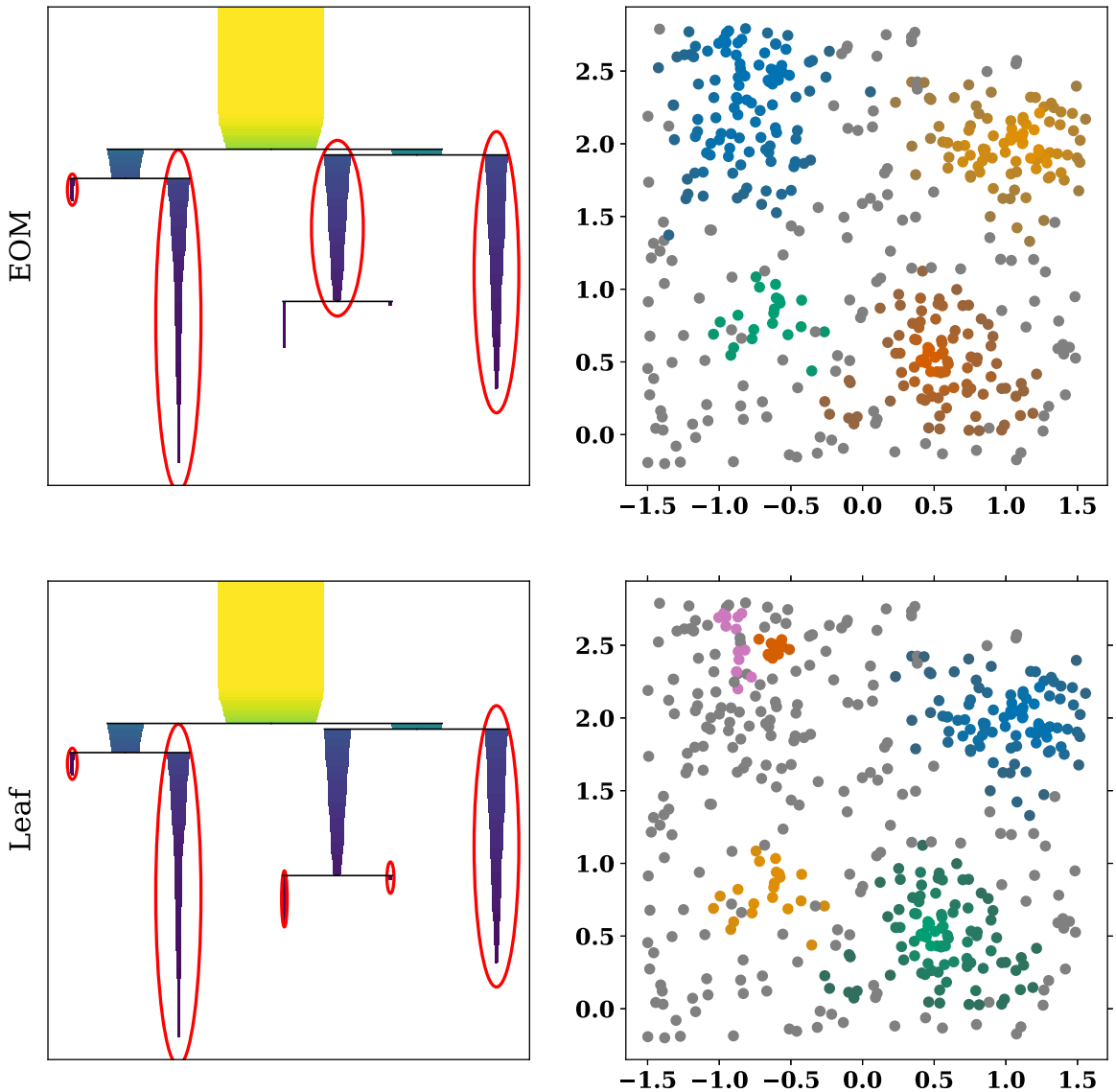

Figure 17. The selection of clusters from the condensed cluster hierarchy with the excess-of-mass (EOM, top row) and leaf (bottom row) methods. Selected branches of the cluster hierarchy are circled in red in the left panels, and the corresponding set of cluster points are shown in color in the right panel, while noise points are gray. The EOM method selects the clusters that persist for the greatest range of density cuts, while the leaf method selects the "leaves" of the cluster hierarchy, i.e., the smallest surviving clusters.

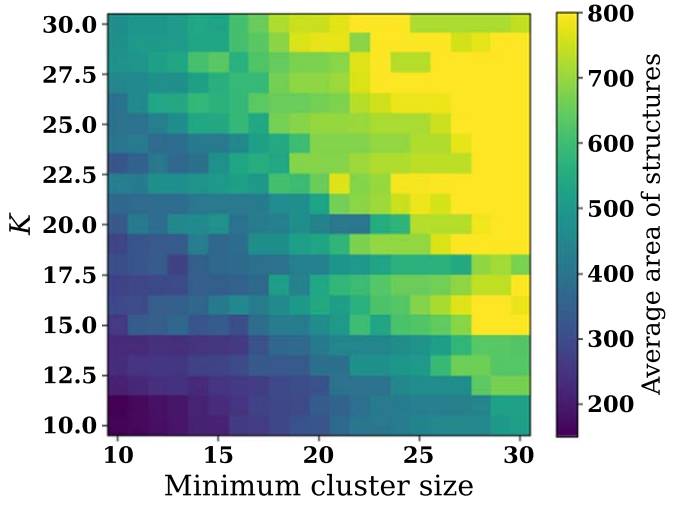

Figure 18. Average area of the structures detected by HDBSCAN, for various combinations of the *K*th nearest neighbor used to measure density (`min_samples`) and minimum number of points to constitute a cluster (`min_cluster_size`). The expected area of a protocluster is ∼100–200 cMpc^2; areas much larger than this suggest that multiple structures are being blended together by the algorithm.

noise, and the cluster selection is more conservative. Setting either of these parameters to too low a value will result in many noise peaks being erroneously selected as clusters, whereas setting them to be too large may result in structures being blended together. Our key consideration in choosing the parameters is ensuring that nearby structures are separated to the extent possible, which we enforce as described below.

We run HDBSCAN on the positions of the LAEs with `min_samples` and `min_cluster_size` both in the range 10–30. We observe that the excess-of-mass method tends to blend structures irrespective of the specific parameter values used; thus, we choose to use the leaf method in our cluster selection. As discussed in the main text, we additionally require that the clusters be above a surface density threshold determined such that the contamination fraction is $\lesssim$0.2. We estimate the mean area of the structures selected for each combination of parameters, as shown in Figure 18. With increasing values of `min_samples` and `min_cluster_size`, the mean area of the detected structures increases, indicating that smaller structures are indeed being blended. We select our final parameters to be `min_cluster_size` = 10 and `min_samples` = 15, which as shown in Figure 18 yields a protocluster sample with a mean area of ∼200 cMpc^2. This choice is based on the physical motivation that the scale of a protocluster is expected to be ∼10–15 cMpc (Y.-K. Chiang et al. 2013), corresponding to a projected area (for an isotropic structure) of ∼100–225 cMpc^2. We make use of a value of `min_samples` on the larger end of the allowed range because, while oversmoothing the density distribution is undesirable, it is equally undesirable for the distribution to be too finely sampled as it becomes very noisy.

Appendix B
Calibration of the Mass Estimation for GS Map and HDBSCAN

In Section 5.2, we evaluated the reliability of the estimated descendant masses for individual protocluster candidates selected from the VT map. We found that the descendant masses were on average underestimated by a factor of ∼3, which we attributed to the underestimation of the protocluster volume by our detection algorithm. Here, we similarly examine the accuracy of the descendant mass estimates for the protocluster candidates selected from the GS map and with HDBSCAN.

We once again use the 90 slices constructed in Section 2.2. Analogously to the middle and right panels of Figure 6, we compare the properties of the structures selected with HDBSCAN (selected from the GS map) from these slices to those of the observed protoclusters in the top (bottom) panel of Figure 19. The structures remain similar in projected area and median LAE overdensity between theory and simulations, reaffirming the fact that our observational protocluster candidates are reasonable.

As in Section 5.2, for each slice, we consider those of the 30 massive cluster progenitors lying within the slice, which are successfully recovered. The recovery rate of the HDBSCAN is ∼70%, slightly higher than that of the VT map. This is in accordance with the fact that the HDBSCAN candidates are the most numerous. By contrast, the recovery rate of the GS map is very low, ∼20%, in line with our observation that only the most prominent structures are recovered.

We compare the estimated and true descendant masses for recovered clusters for both methods in Figure 20. We find that the descendant masses of the protoclusters are underestimated by a factor of ∼2 with the GS method, and by a factor of ∼1.15 with HDBSCAN. In both cases, the scatter is similar to that observed with the VT map, ∼0.5 dex.

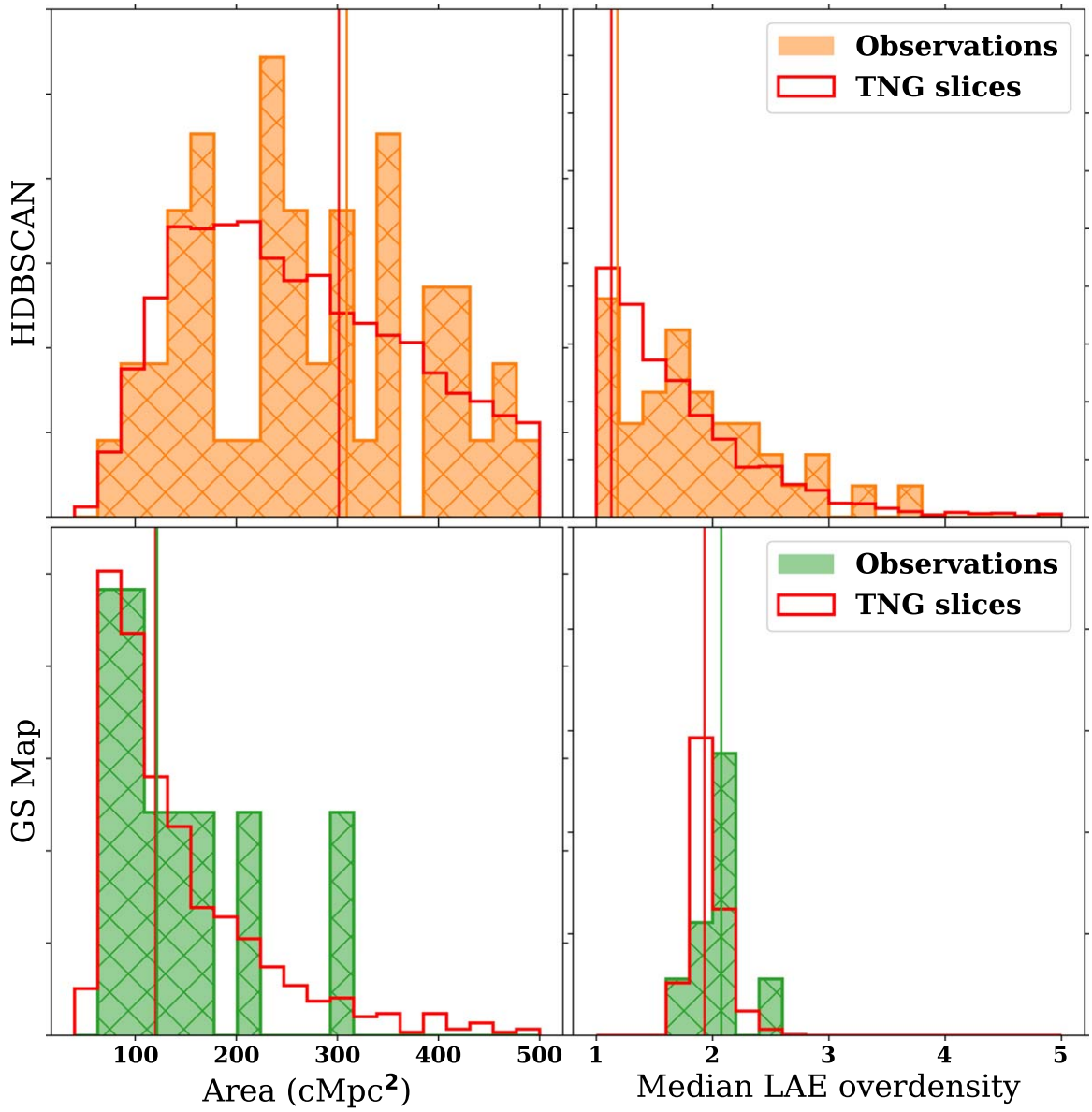

Figure 19. Area and median LAE overdensity ($\delta_{\rm LAE}$) for the structures selected with HDBSCAN (top) and from the GS map (bottom) in the TNG300 slices compared to those in observations. Solid lines indicate the median values of each property. As with the structures selected from the VT map, the properties of the objects in the simulation and observations are similar.

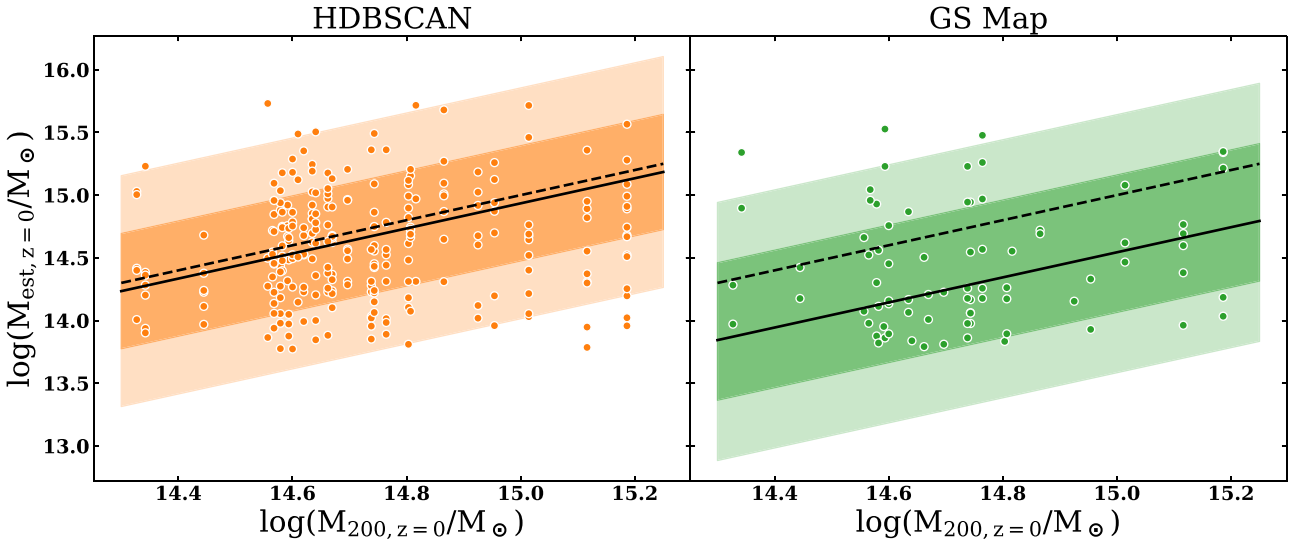

Figure 20. Descendant mass estimated according to Equation (6) compared to the true descendant mass for structures selected with HDBSCAN (left) and from the GS map (right). The dashed lines represent a 1-to-1 relation of estimated and true mass whereas the solid lines represent the median ratio of the two.

ORCID iDs

Vandana Ramakrishnan https://orcid.org/0000-0002-9176-7252
Kyoung-Soo Lee https://orcid.org/0000-0003-3004-9596
Maria Celeste Artale https://orcid.org/0000-0003-0570-785X
Eric Gawiser https://orcid.org/0000-0003-1530-8713
Yujin Yang https://orcid.org/0000-0003-3078-2763
Changbom Park https://orcid.org/0000-0001-9521-6397
Yi-Kuan Chiang https://orcid.org/0000-0001-6320-261X
Robin Ciardullo https://orcid.org/0000-0002-1328-0211
Arjun Dey https://orcid.org/0000-0002-4928-4003
Caryl Gronwall https://orcid.org/0000-0001-6842-2371
Lucia Guaita https://orcid.org/0000-0002-4902-0075
Ho Seong Hwang https://orcid.org/0000-0003-3428-7612
Sang Hyeok Im https://orcid.org/0009-0003-9748-4194
Woong-Seob Jeong https://orcid.org/0000-0002-2770-808X
Seongjae Kim https://orcid.org/0009-0002-3931-6697
Ankit Kumar https://orcid.org/0000-0001-6270-3527
Jaehyun Lee https://orcid.org/0000-0002-6810-1778
Seong-Kook Lee https://orcid.org/0000-0001-5342-8906
Byeongha Moon https://orcid.org/0009-0008-4022-3870
Nelson Padilla https://orcid.org/0000-0001-9850-9419
Alexandra Pope https://orcid.org/0000-0001-8592-2706
Roxana Popescu https://orcid.org/0000-0001-8245-7669
Akriti Singh https://orcid.org/0009-0000-7651-3900
Hyunmi Song https://orcid.org/0000-0002-4362-4070
Paulina Troncoso https://orcid.org/0000-0001-6162-3023
Francisco Valdes https://orcid.org/0000-0001-5567-1301
Ann Zabludoff https://orcid.org/0000-0001-6047-8469

References

Aihara, H., AlSayyad, Y., Ando, M., et al. 2019, PASJ, 71, 114
Aihara, H., Arimoto, N., Armstrong, R., et al. 2018a, PASJ, 70, S4
Aihara, H., Armstrong, R., Bickerton, S., et al. 2018b, PASJ, 70, S8
Alpaslan, M., Robotham, A. S. G., Driver, S., et al. 2014, MNRAS, 438, 177
Andrews, M., Artale, M. C., Kumar, A., et al. 2024, arXiv:2410.08412
Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., et al. 2022, ApJ, 935, 167
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123
Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33
Barbary, K. 2016, JOSS, 1, 58
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Bielby, R. M., Tummuangpak, P., Shanks, T., et al. 2016, MNRAS, 456, 4061
Bleem, L. E., Stalder, B., de Haan, T., et al. 2015, ApJS, 216, 27
Bond, J. R., Kofman, L., & Pogosyan, D. 1996, Natur, 380, 603
Boylan-Kolchin, M., Springel, V., White, S. D. M., Jenkins, A., & Lemson, G. 2009, MNRAS, 398, 1150
Bădescu, T., Yang, Y., Bertoldi, F., et al. 2017, ApJ, 845, 172
Byrohl, C., & Nelson, D. 2023, MNRAS, 523, 5248
Campello, R. J. G. B., Moulavi, D., & Sander, J. 2013, in Advances in Knowledge Discovery and Data Mining, ed. J. Pei et al. (Berlin: Springer), 160
Casey, C. M., Cooray, A., Capak, P., et al. 2015, ApJL, 808, L33
Cautun, M. C., & van de Weygaert, R. 2011, arXiv:1105.0370
Chiang, Y.-K., Overzier, R., & Gebhardt, K. 2013, ApJ, 779, 127
Chiang, Y.-K., Overzier, R., & Gebhardt, K. 2014, ApJL, 782, L3
Ciardullo, R., Feldmeier, J. J., Krelove, K., Jacoby, G. H., & Gronwall, C. 2002, ApJ, 566, 784
Ciardullo, R., Gronwall, C., Wolf, C., et al. 2012, ApJ, 744, 110
Cucciati, O., Lemaux, B. C., Zamorani, G., et al. 2018, A&A, 619, A49
Daddi, E., Valentino, F., Rich, R. M., et al. 2021, A&A, 649, A78
Darvish, B., Mobasher, B., Sobral, D., Scoville, N., & Aragon-Calvo, M. 2015, ApJ, 805, 121
Dekel, A., Birnboim, Y., Engel, G., et al. 2009, Natur, 457, 451
Dey, A., Lee, K.-S., Reddy, N., et al. 2016, ApJ, 823, 11
Dijkstra, M., & Westra, E. 2010, MNRAS, 401, 2343
Eisenhardt, P. R. M., Brodwin, M., Gonzalez, A. H., et al. 2008, ApJ, 684, 905
Firestone, N. M., Gawiser, E., Ramakrishnan, V., et al. 2024, ApJ, 974, 217
Gawiser, E., Francke, H., Lai, K., et al. 2007, ApJ, 671, 278
Gebhardt, K., Mentuch Cooper, E., Ciardullo, R., et al. 2021, ApJ, 923, 217
Gonzalez, A. H., Gettings, D. P., Brodwin, M., et al. 2019, ApJS, 240, 33
Gronwall, C., Ciardullo, R., Hickey, T., et al. 2007, ApJ, 667, 79
Guaita, L., Gawiser, E., Padilla, N., et al. 2010, ApJ, 714, 255
Hagen, A., Ciardullo, R., Gronwall, C., et al. 2014, ApJ, 786, 59
Harikane, Y., Ouchi, M., Ono, Y., et al. 2019, ApJ, 883, 142
Hayashi, M., Tanaka, M., Shimakawa, R., et al. 2018, PASJ, 70, S17
Hayashino, T., Matsuda, Y., Tamura, H., et al. 2004, AJ, 128, 2073
Hopkins, P. F., Hernquist, L., Cox, T. J., et al. 2006, ApJS, 163, 1
Huang, Y., Lee, K.-S., Cucciati, O., et al. 2022, ApJ, 941, 134
Hung, D., Lemaux, B. C., Gal, R. R., et al. 2020, MNRAS, 491, 5524
Im, S. H., Hwang, H. S., Park, J., et al. 2024, ApJ, 972, 196
Ito, K., Tanaka, M., Valentino, F., et al. 2023, ApJL, 945, L9
Kikuta, S., Matsuda, Y., Cen, R., et al. 2019, PASJ, 71, L2
Koyama, Y., Kodama, T., Tadaki, K.-i., et al. 2014, ApJ, 789, 18
Kraljic, K., Arnouts, S., Pichon, C., et al. 2017, MNRAS, 474, 547
Kriegel, H.-P., Kröger, P., Sander, J., & Zimek, A. 2011, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1, 231

Kuchner, U., Haggar, R., Aragón-Salamanca, A., et al. 2022, MNRAS, 510, 581
Kusakabe, H., Shimasaku, K., Ouchi, M., et al. 2018, PASJ, 70, 4
Laigle, C., Pichon, C., Arnouts, S., et al. 2018, MNRAS, 474, 5437
Lee, K.-S., Dey, A., Hong, S., et al. 2014, ApJ, 796, 126
Lee, K.-S., Gawiser, E., Park, C., et al. 2024, ApJ, 962, 36
Lemaux, B. C., Cucciati, O., Le Fèvre, O., et al. 2022, A&A, 662, A33
Lemaux, B. C., Le Fèvre, O., Cucciati, O., et al. 2018, A&A, 615, A77
Madau, P., & Dickinson, M. 2014, ARA&A, 52, 415
Malavasi, N., Aghanim, N., Douspis, M., Tanimura, H., & Bonjean, V. 2020, A&A, 642, A19
Malavasi, N., Arnouts, S., Vibert, D., et al. 2017, MNRAS, 465, 3817
Malavasi, N., Lee, K.-S., Dey, A., et al. 2021, ApJ, 921, 103
Matsuda, Y., Nakamura, Y., Morimoto, N., et al. 2009, MNRAS, 400, L66
Matsuda, Y., Yamada, T., Hayashino, T., et al. 2004, AJ, 128, 569
Matthee, J., Sobral, D., Best, P., et al. 2017, MNRAS, 471, 629
McInnes, L., Healy, J., & Astels, S. 2017, JOSS, 2, 205
Momose, R., Shimasaku, K., Kashikawa, N., et al. 2021, ApJ, 909, 117
Nelson, D., Springel, V., Pillepich, A., et al. 2019, ComAC, 6, 2
Oteo, I., Ivison, R. J., Dunne, L., et al. 2018, ApJ, 856, 72
Ouchi, M., Shimasaku, K., Akiyama, M., et al. 2008, ApJS, 176, 301
Overzier, R. A. 2016, A&ARv, 24, 14
Peng, Y.-j., Lilly, S. J., Kovač, K., et al. 2010, ApJ, 721, 193
Pillepich, A., Nelson, D., Hernquist, L., et al. 2018a, MNRAS, 475, 648
Pillepich, A., Springel, V., Nelson, D., et al. 2018b, MNRAS, 473, 4077
Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016, A&A, 594, A13
Quadri, R. F., Williams, R. J., Franx, M., & Hildebrandt, H. 2012, ApJ, 744, 88
Ramakrishnan, V., Moon, B., Im, S. H., et al. 2023, ApJ, 951, 119
Ravi, J., Hadzhiyska, B., White, M., Hernquist, L., & Bose, S. 2024, PhRvD, 110, 103509
Reddy, N. A., & Steidel, C. C. 2009, ApJ, 692, 778
Rykoff, E. S., Rozo, E., Busha, M. T., et al. 2014, ApJ, 785, 104
Rykoff, E. S., Rozo, E., Hollowood, D., et al. 2016, ApJS, 224, 1
Salerno, J. M., Martínez, H. J., & Muriel, J. M. 2019, MNRAS, 484, 2
Sarron, F., Adami, C., Durret, F., & Laigle, C. 2019, A&A, 632, A49
Sarron, F., & Conselice, C. J. 2021, MNRAS, 506, 2136
Sawicki, M., Arnouts, S., Huang, J., et al. 2019, MNRAS, 489, 5202
Schaap, W. E., & van de Weygaert, R. 2000, A&A, 363, L29
Shi, K., Huang, Y., Lee, K.-S., et al. 2019, ApJ, 879, 9
Shi, K., Toshikawa, J., Lee, K.-S., et al. 2021, ApJ, 911, 46
Shimakawa, R., Kodama, T., Hayashi, M., et al. 2017, MNRAS, 468, L21
Sousbie, T. 2011, MNRAS, 414, 350
Sousbie, T., Pichon, C., & Kawahara, H. 2011, MNRAS, 414, 384
Springel, V., White, S. D. M., Jenkins, A., et al. 2005, Natur, 435, 629
Stark, D. P., Ellis, R. S., Chiu, K., Ouchi, M., & Bunker, A. 2010, MNRAS, 408, 1628
Steidel, C. C., Adelberger, K. L., Dickinson, M., et al. 1998, ApJ, 492, 428
Steidel, C. C., Adelberger, K. L., Shapley, A. E., et al. 2000, ApJ, 532, 170
Tempel, E., Stoica, R. S., Martínez, V. J., et al. 2014, MNRAS, 438, 3465
Toshikawa, J., Kashikawa, N., Overzier, R., et al. 2016, ApJ, 826, 114
Toshikawa, J., Uchiyama, H., Kashikawa, N., et al. 2018, PASJ, 70, S12
Umehata, H., Tamura, Y., Kohno, K., et al. 2015, ApJL, 815, L8
van de Weygaert, R., & Schaap, W. 2009, in Data Analysis in Cosmology, ed. V. J. Martínez et al., Vol. 665 (Berlin: Springer), 291
van der Burg, R. F. J., Muzzin, A., Hoekstra, H., et al. 2013, A&A, 557, A15
Vargas, C. J., Bish, H., Acquaviva, V., et al. 2014, ApJ, 783, 26
Wang, T., Elbaz, D., Daddi, E., et al. 2016, ApJ, 828, 56
Weiss, L. H., Bowman, W. P., Ciardullo, R., et al. 2021, ApJ, 912, 100
White, M., Raichoor, A., Dey, A., et al. 2024, JCAP, 2024, 020
Yamada, T., Nakamura, Y., Matsuda, Y., et al. 2012, AJ, 143, 79
Zhu, W., Zhang, F., & Feng, L.-L. 2021, ApJ, 920, 2