

UC Irvine

UC Irvine Previously Published Works

Title

Galaxy-multiplet clustering from DESI DR2

Permalink

<https://escholarship.org/uc/item/90j444fn>

Journal

Monthly Notices of the Royal Astronomical Society, 545(4)

ISSN

0035-8711

Authors

Wang, Hanyue
Eisenstein, Daniel J
Aguilar, Jessica Nicole
[et al.](#)

Publication Date

2026-01-05

DOI

10.1093/mnras/staf2069

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Galaxy-multiplet clustering from DESI DR2

Hanyue Wang^{1,2,3}★ Daniel J. Eisenstein¹ Jessica Nicole Aguilar⁴ Steven Ahlen⁵ Davide Bianchi^{6,7} David Brooks⁸ Todd Claybaugh⁴ Axel de la Macorra⁹ Arjun Dey¹⁰ Biprateep Dey^{11,12} Peter Doel⁸ Simone Ferraro^{4,13} Andreu Font-Ribera¹⁴ Jaime E. Forero-Romero^{15,16} Enrique Gaztañaga^{17,18,19} Gaston Gutierrez²⁰ Klaus Honscheid^{21,22,23} Mustapha Ishak²⁴ Richard Joyce¹⁰ Stephanie Juneau¹⁰ David Kirkby²⁵ Theodore Kisner⁴ Anthony Kremin⁴ Ofer Lahav⁸ Claire Lamman²³ Martin Landriau⁴ Marc Manera^{14,26} Aaron Meisner¹⁰ Ramon Miquel^{14,27} Eva-Maria Mueller²⁸ Seshadri Nadathur¹⁸ Gustavo Niz^{29,30} Nathalie Palanque-Delabrouille^{4,31} Will J. Percival^{32,33,34} Francisco Prada³⁵ Ignasi Pérez-Ràfols³⁶ Ashley J. Ross^{21,23,37} Graziano Rossi³⁸ Eusebio Sanchez³⁹ David Schlegel⁴ Michael Schubnell^{40,41} Joseph Harry Silber⁴ David Sprayberry¹⁰ Gregory Tarlé⁴¹ Benjamin Alan Weaver¹⁰ Rongpu Zhou⁴ and Hu Zou⁴²

Affiliations are listed at the end of the paper

Accepted 2025 November 16. Received 2025 November 13; in original form 2025 September 19

ABSTRACT

We present an efficient estimator for higher order galaxy clustering using small groups of nearby galaxies, or multiplets. Using the Luminous Red Galaxy sample from the Dark Energy Spectroscopic Instrument (DESI) Data Release 2, we identify galaxy multiplets as discrete objects and measure their cross-correlations with the general galaxy field. Our results show that the multiplets exhibit stronger clustering bias as they trace more massive dark matter haloes than individual galaxies. When comparing the observed clustering statistics with the mock catalogues generated from the N -body simulation ABACUSSUMMIT, we find that the mocks underpredict multiplet clustering despite reproducing the galaxy two-point autocorrelation reasonably well. This discrepancy indicates that the standard Halo Occupation Distribution (HOD) model is insufficient to describe the properties of galaxy multiplets, revealing the greater constraining power of this higher order statistic on galaxy–halo connection and the possibility that multiplets are specific to additional assembly bias. We demonstrate that incorporating secondary biases into the HOD model improves agreement with the observed multiplet statistics, specifically by allowing galaxies to preferentially occupy haloes in denser environments. Our results highlight the potential of utilizing multiplet clustering, beyond traditional two-point correlation measurements, to break degeneracies in models describing the galaxy–dark matter connection.

Key words: methods: data analysis – large-scale structure of Universe – cosmology: observations.

1 INTRODUCTION

Galaxy clustering encodes valuable information about large-scale structure (LSS) and cosmology. Since galaxies are biased tracers of the underlying dark-matter density field, measurements of their clustering properties provide critical constraints on both cosmological parameters and the physics of galaxy formation, specifically the relationship between galaxies and dark matter haloes (see R. H. Wechsler & J. L. Tinker 2018 for a review). With the advent of advanced sky surveys, we are entering an era of high-precision studies of galaxy clustering, where galaxy distribution needs to be effectively quantified to capture the full complexity of galaxy bias and deepen our understanding of cosmology and galaxy–dark matter connection.

The most commonly used statistical tool to quantify galaxy clustering is the two-point correlation function (2PCF; P. J. E. Peebles 1980; M. Davis & P. J. E. Peebles 1983), which should be sufficient to describe the spatial matter distribution of the universe if the matter density fluctuations constitute a Gaussian random field (J. M. Bardeen et al. 1986). However, non-linear gravitational instabilities and the complex physics of galaxy formation produce non-Gaussian signatures in the galaxy distribution measured today (F. Bernardeau et al. 2002). Therefore, higher order statistics are necessary to complement the information provided by the 2PCF when studying late-time galaxy distribution. Measurements of higher order statistics using galaxy catalogues have a long-standing history. For example, there has been a wealth of studies using the three-point functions (3PCF) to constrain the galaxy–halo connection and to break degeneracies between cosmology and galaxy bias with both simulations and observations (E. Gaztañaga et al. 2005; G. V. Kulkarni et al. 2007; F. Marín 2011; C. K. McBride et al. 2011; Z.

* E-mail: hanyuew@stanford.edu

Slepian et al. 2017; S. Yuan, D. J. Eisenstein & L. H. Garrison 2018; H. Zhang et al. 2022). However, it is computationally expensive to calculate higher order correlation functions, especially with a growing sample size from current and upcoming galaxy surveys.

Therefore, in this work, we use galaxy multiplets, small groups of nearby galaxies, to study higher order clustering by cross-correlating them with the full galaxy sample. Unlike galaxy groups or clusters, multiplets are not necessarily virialized, but their physical configurations occupy the regime of gravitational non-linearities and thus contain rich information. We expect multiplets to trace dark matter differently than individual galaxies, which can be probed through their clustering patterns. In practice, we identify galaxy multiplets based on their spatial information, treat multiplets of n galaxies as discrete objects, and measure the cross-correlation between the multiplet field and the general galaxy field. This cross-correlation is related to an $n + 1$ point correlation function in the squeezed limit, referring to the configuration in which one of the vertices is significantly more separated from the rest. Through this cross-correlation between the multiplet field and the general galaxy field, the computationally expensive multipoint calculation is reduced to two separate pair-finding problems while preserving the information from higher order correlations and small-scale non-Gaussianity.

This approach extends the idea of the squeezed 3PCF framework introduced in S. Yuan, D. J. Eisenstein & L. H. Garrison (2017), which focused on close galaxy pairs, with several modifications. As S. Yuan et al. (2017) use all available galaxy pairs to compute the cross-correlation, the measurement is effectively equivalent to a standard squeezed 3PCF. In this work, however, we explicitly separate the multiplets into systems of distinct size (i.e. pairs, triplets, and quads), so that subgroups of higher multiplicity systems do not enter lower order analyses. For example, a triplet can be viewed as three different pairs, but they are not included in the galaxy-pair cross-correlation calculations. Consequently, our measurement differs from the standard framework of squeezed higher order correlations, but it remains closely related and retains higher order information in the multiplet configurations. We choose this separation because systems of different multiplicities are physically distinct tracers of dark matter, and measuring the cross-correlation functions separately provides access to their physical properties directly.

Multiplets are physically interesting systems that can probe various aspects of the LSS. While multiplet orientations have been used in the perspective of intrinsic alignment to study the large-scale tidal field (C. Lamman et al. 2024), here we focus on their clustering properties. The clustering of multiplets is directly connected to the underlying matter density field, providing key information about non-linear structure formation and the galaxy-halo connection. While reconstructing galaxy formation through the combination of large-scale matter distribution and galaxy bias remains uncertain, the multipoint correlation functions measured with multiplets will likely probe how galaxies trace dark matter in novel ways.

We explore this new estimator of multiplet clustering using the Luminous Red Galaxy (LRG) sample from the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration 2016a, b, 2022; J. H. Silber et al. 2023; T. N. Miller et al. 2024; C. Poppett et al. 2024). We develop an efficient algorithm to identify galaxy multiplets (specifically pairs, triplets, and quads) and measure the galaxy-multiplet cross-correlation functions by cross-correlating the multiplet field with the general galaxy field. To evaluate the effectiveness of this higher order statistic, we compare the measured multiplet clustering with predictions from mock catalogues, demonstrating its potential to probe assembly bias and to place tighter constraints on the complex galaxy-dark matter connection.

This paper is structured as follows. In Section 2, we describe the data sample used for the study and outline the identification of galaxy multiplets. In Section 3, we introduce the correlation methodology for measuring multiplet clustering and present the corresponding results. We introduce the DESI mock catalogues in Section 4 and discuss the calibration for survey systematics in Section 5. In Section 6, we compare the clustering measurements from the mocks to observations and explore the implications for the galaxy-halo connection. Finally, we summarize and discuss the results further in Section 7. Throughout the paper, we assume $H_0 = 70$ km/s/Mpc.

2 DATA AND MULTIPLET IDENTIFICATION

2.1 The DESI LRG sample

DESI started the main survey observation in 2021 May after a period of survey validation (DESI Collaboration 2024a). This study makes use of the DESI Year 3 data set, collected through 2024 April, with processing via the ‘Loa’ version of the data compilation using the DESI spectroscopic reduction pipeline (J. Guy et al. 2023) and the redshift estimation pipeline (REDROCK, A. Anand et al. 2024; S. Bailey et al. in preparation). It will be released as part of the DESI Data Release 2 (DR2) in a manner similar to the Early Data Release and the Data Release 1 (DR1, DESI Collaboration 2025e).

LRGs are an important class of galaxies for LSS studies and are specifically selected in this work for two reasons. First, the sample is relatively complete due to their relatively high priority in the DESI fibre assignment procedure; and secondly, they are highly biased tracers of the LSS, offering opportunities to probe high-density environments.

LRG target selection (R. Zhou et al. 2023) is based on Legacy Imaging Surveys Data Release 9 (H. Zou et al. 2017; A. Dey et al. 2019), specifically the g, r, z optical bands, and the *Wide-field Infrared Survey Explorer* (E. L. Wright et al. 2010) photometry. Additional information on DESI’s observation planning and target selection is available in E. F. Schlafly et al. (2023) and A. D. Myers et al. (2023). For the clustering analysis, we use LRGs from the DR2 LSS catalogue (A. J. Ross et al. 2025), which incorporates details of target selection. Compared to DR1, the fibre-assignment completeness has increased by 19 per cent (DESI Collaboration 2025a), which is important for our study, as missing galaxies in the identified galaxy multiplets due to fibre-assignment incompleteness can bias the calculation of multiplet cross-correlations.

Fig. 1 shows the footprint of the LRGs colour-coded by NTILE, which describes the number of times a certain portion of the sky has been observed. We choose to focus on a region with a relatively large value of NTILE (and then high completeness) to avoid missing galaxies in the identified multiplets and thus to reduce measurement bias in clustering statistics. This selected region, which covers approximately 5700 deg^2 with a fibre assignment completeness of approximately 90%, is marked with the red contour in Fig. 1.

2.2 Multiplet-identification algorithm

The galaxy multiplets are identified using their spatial information. For each galaxy, we calculate its position on the Cartesian coordinate based on its angular coordinates RA and Dec., and its spectroscopic redshift z . We utilize a flat cosmological model with matter density $\Omega_m = 0.3$ and cosmological constant density $\Omega_\Lambda = 0.7$ to compute the comoving distance using the redshift. With the Cartesian coordinates, we look for close galaxy pairs with a separation less than 10 Mpc and then further select the pairs based on a transverse

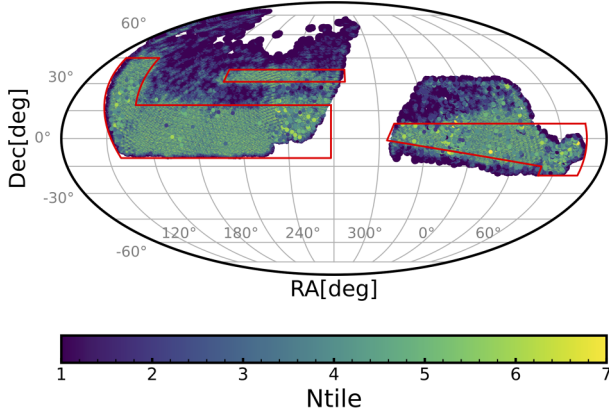


Figure 1. The footprint of the LRGs from DESI DR2, colour-coded by NTILE. The solid contour marks the region with relatively high completeness that our study focuses on.

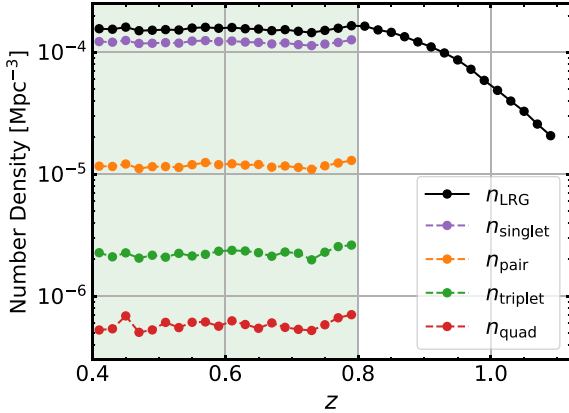


Figure 2. The redshift distributions of the LRGs and the multiplets used as tracers in the clustering analysis. The shaded area marks the selected redshift range for this work, where the number densities of all tracers are approximately constant.

separation threshold of 1.5 Mpc. The line-of-sight (LOS) separation is chosen to be much larger than the transverse separation to account for redshift–space distortions, especially the ‘Fingers of God’ effect. We calibrate the transverse separation as the length of the transverse separation vector, which is calculated by subtracting the dot product of the separation vector between the galaxy pair and the vector pointing to their midpoint from the separation vector. After obtaining the pairs, we utilize a friends-of-friends algorithm to link the pairs together by assigning the same group number to all pairs containing the same galaxy, forming the catalogue of galaxy multiplets. The remaining galaxies that are not part of any larger multiplet are classified as singlets.

Fig. 2 shows the redshift distributions of the LRGs and the multiplets (singlets, pairs, triplets, and quads) used as tracers in this work. At $z > 0.8$, the LRG density drops rapidly due to the flux limit of the sample in the imaging. To focus on the regime where the number density remains approximately constant, we restrict the analysis to LRGs in the redshift range of $0.4 < z < 0.8$, which also corresponds to the first two LRG bins in the DESI 2024 cosmological analysis (DESI Collaboration 2025c, c, 2025d). The redshift distribution of the multiplets demonstrates that there is

little redshift evolution within this range for all tracers used in the clustering analysis.

To summarize some statistics, from the total 1603 265 LRGs in a sky area of roughly 5700 deg^2 , we find 230 863 galaxy pairs, which combine into 155 267 multiplets of minimum size two using the friends-of-friends method. Among all the LRGs, about 22.3 per cent of galaxies reside in larger multiplets, while the others remain as singlets; among all the galaxies grouped as multiplets, about 68.6 per cent of them appear in pairs. In total, we identify 122 588 pairs, 23 508 triplets, and 6092 quads. It is worth noting that as we approach larger galaxy multiplets, their number densities decrease significantly.

3 CORRELATION METHODOLOGY AND RESULTS

3.1 Clustering statistics

In this section, we present the analytical tools used in our work, including the pair-count-based estimators for the autocorrelation and the galaxy-multiplet cross-correlation functions, together with the method for error estimation.

The calculation of correlation functions requires ‘random’ catalogues, which contain samples of random sky positions that directly represent the survey data in sky geometry and redshift distribution. High-quality random catalogues are essential for estimators to properly represent the correlation function, as the measurements depend sensitively on the randoms to accurately reflect spatial and redshift-dependent selection effects in the actual survey data. DESI randoms are processed through the same pipeline as the LSS catalogues and are standardized to a number density of 2500 deg^{-2} .

To ensure that clustering measurements are unbiased from survey specifics, each object in the DESI LSS and random catalogues is assigned a weight, as detailed in DESI Collaboration (2025c). Briefly, each observed target in the data catalogue is assigned a total weight to account for the DESI selection function. This total weight is a product of three components: the completeness weight `WEIGHT_COMP`, the redshift failure weight `WEIGHT_ZFAIL`, and the systematic weight `WEIGHT_SYS`. Specifically, `WEIGHT_COMP` describes how well sampled the targets are in a given region of the sky. It is primarily correlated with the NTILE parameter and the geometry of DESI’s fibre positioners on the focal plane, accounting for fibre assignment incompleteness and fibre collisions (DESI Collaboration 2024b). Then, `WEIGHT_SYS` corrects for variations in the imaging data quality, and `WEIGHT_ZFAIL` accounts for the variations in accuracy and precision of redshift estimation in particular DESI observations. Weights for objects in the random catalogues are assigned from randomly chosen galaxies in the data catalogues, along with their associated redshifts, to obtain the same redshift distribution as the data. Ultimately, the estimator for the correlation function adopts the normalized weighted pair counts, with each pair weighted by the product of the individual weights assigned to the two objects involved.

With the DESI LSS and random catalogues, we start with the galaxy–galaxy autocorrelation by computing the anisotropic 2PCF $\xi(r_p, \pi)$ with the widely used Landy–Szalay estimator (S. D. Landy & A. S. Szalay 1993):

$$\xi(r_p, \pi) = \frac{DD(r_p, \pi) - 2DR(r_p, \pi) + RR(r_p, \pi)}{RR(r_p, \pi)}, \quad (1)$$

where r_p is the transverse distance and π is the distance along the LOS. $DD(r_p, \pi)$, $RR(r_p, \pi)$, and $DR(r_p, \pi)$ are the normalized

weighted pair counts between galaxy–galaxy, random–random, and galaxy–random catalogues in each particular (r_p, π) bin. Specifically, we use 14 r_p bins spanning the interval $[0.01, 30]$ Mpc with a maximum LOS separation of $\pi_{\max} = 60$ Mpc.

To compute the galaxy–multiplet cross-correlations, we treat multiplets of different sizes as discrete objects at the average angular coordinates and redshift of their member galaxies, and cross-correlate them with the full galaxy sample. Because it is non-trivial to construct random catalogues for the multiplets, we adopt the asymmetric Davis–Peebles estimator:

$$\xi(r_p, \pi) = \frac{D_m D_g(r_p, \pi)}{D_m R_g(r_p, \pi)} - 1, \quad (2)$$

where D_m denotes the data catalogue of multiplets (i.e. pairs, triples, and quads), and D_g and R_g correspond to the data and random catalogues for the entire selected DESI LRG sample. Each object in D_g and R_g is weighted in the same way as in the autocorrelation calculation, following the standard DESI procedure described above. All multiplets in D_m are assigned a weight of 1, since their multiplicity introduces more complexity in the weighting due to biases from incompleteness and other systematics.

The projected correlation function is then obtained by integrating the anisotropic 2D correlation function $\xi(r_p, \pi)$ along the LOS,

$$w_p(r_p) = \int_{-\pi_{\max}}^{\pi_{\max}} \xi(r_p, \pi) d\pi. \quad (3)$$

As the projected correlation marginalizes over the direction along the LOS, it is less informative than the full-shape $\xi(r_p, \pi)$. Still, we choose to use w_p in our main analysis because it is easy to visualize as a 1D function and avoids the complexity of modelling galaxy velocities.

All pair-counting calculations in this work are performed with the python package `Corrfunc` (M. Sinha & L. H. Garrison 2020).

Finally, to estimate the uncertainties in the projected correlations, we divide the catalogues of galaxy multiplets into 50 spatial subsamples but keep the D_g and R_g catalogues in their original form, repeat the calculations for w_p while dropping one multiplet subsample at a time, and then compute the error bars with jackknife error estimation:

$$\text{Var}(x) = \frac{n-1}{n} \sum_{i=1}^n (\bar{x}_i - \bar{x}_{\text{jack}})^2, \quad \bar{x}_{\text{jack}} = \frac{1}{n} \sum_{i=1}^n \bar{x}_i, \quad (4)$$

where $\bar{x}_i$ is the i th jackknife replicate calculated by dropping the i th multiplet subsample, and $n = 50$ is the total number of jackknife regions.

3.2 Results

Following the methodology, we present the clustering measurements from the LRGs. Fig. 3 shows the galaxy–galaxy autocorrelation, along with the cross-correlations between galaxy singlets/multiplets (specifically pairs, triplets, and quads) and the general galaxy field, with the error bars estimated using the jackknife method. Singlets here refer to galaxies that do not belong to any larger multiplet, so their cross-correlation with the full galaxy sample is expected to be lower than the two-point autocorrelation. Meanwhile, the multiplets exhibit stronger clustering bias as they trace more massive dark matter haloes. To ensure reliable statistics and spatial coverage, we require the multiplets to be sufficiently abundant across the analysis field. For instance, the cross-correlation with quads shows more fluctuations on large scales ($10 \text{ Mpc} < r_p < 30 \text{ Mpc}$) compared to those with smaller multiplets, reflecting their sensitivity to cosmic

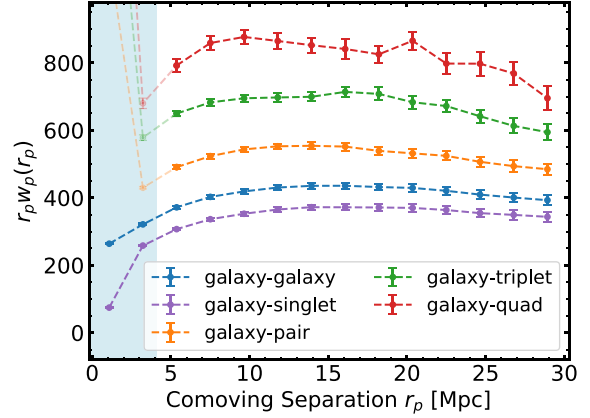


Figure 3. Comparison between the projected functions for the galaxy two-point autocorrelation and the cross-correlations with singlets, pairs, triplets, and quads.

variance. Therefore, we limit the multiplet analysis to systems with up to four galaxies.

On small scales, specifically $r_p \lesssim 2$ Mpc, we note that the cross-correlations with galaxy multiplets increase sharply. This is because our multiplet identification algorithm sets the threshold for galaxy pairs to be 1.5 Mpc along the projected direction, which leads to these overclustering signatures observed at the scale of individual galaxy multiplet. Meanwhile, the galaxy-singlet cross-correlation dips further below the autocorrelation because most small-scale clustering signals arise from the correlation between satellite galaxies and their hosts (the one-halo term in the literature), which have been largely excluded from the singlet selection. As a result, we exclude measurements for the multiplets on small scales ($r_p < 4$ Mpc, shown as the shaded area in Fig. 3), but only retain small-scale information of the autocorrelation.

In general, multiplets trace the underlying dark matter field differently from individual galaxies. We observe excess clustering for higher order multiplets, consistent with their association with more massive dark matter haloes. Relative to the autocorrelation, the cross-correlation with singlets has approximately 0.86 of the amplitude, while the cross-correlations with pairs, triplets, and quads are boosted by factors of ~ 1.25 , 1.60, and 1.94, respectively, demonstrating that larger multiplets are more biased tracers of the dark matter density field. These statistics will be valuable for comparison with the clustering properties constructed from simulations in the following sections.

4 SIMULATIONS

To connect observations with theoretical predictions, it is necessary to model the underlying dark matter density field of the universe and construct mock galaxy catalogues on top of it following some galaxy–dark matter connection model. In this section, we describe the construction of the official DESI mocks and outline the key components used to generate the mock galaxy catalogues for our analysis.

4.1 AbacusSummit

The DESI mocks are built with the dark matter density fields from the ABACUSUMMIT simulation suite, a set of large cosmological simulations using the high-performance high-accuracy ABACUS N -

body code (L. H. Garrison et al. 2021; N. A. Maksimova et al. 2021). These simulations were specifically developed to satisfy the Cosmological Simulation Requirements of DESI. ABACUSSUMMIT includes 25 base simulations at Planck 2018 Lambda cold dark matter (Λ CDM) cosmology (Planck Collaboration VI 2020), each containing $6912^3 \approx 330$ billion particles with a particle mass of $2 \times 10^9 h^{-1} M_\odot$ in a periodic box of volume $(2 h^{-1} \text{Gpc})^3$.¹

4.2 Halo Occupation Distribution

To connect the galaxy distribution to the dark matter density field and to generate mock galaxy catalogues using halo catalogues from ABACUSSUMMIT, we employ the important framework of Halo Occupation Distribution (HOD). HOD models describe how galaxies populate dark matter haloes as a function of different halo properties (A. Cooray & R. Sheth 2002). Statistically, the HOD formalism models the conditional probability $P(n|\mathbf{X}_h)$ that a halo with properties $\mathbf{X}_h$ hosts n galaxies of a given type. In addition to predicting galaxy counts, HOD models also characterize the spatial and velocity distributions of galaxies relative to dark matter within haloes.

In its simplest implementation, the vanilla HOD model distinguishes between central and satellite galaxies based on their different formation mechanisms and assumes that halo mass is the only relevant parameter among all halo properties when assigning galaxies to haloes (Z. Zheng et al. 2005). Meanwhile, extensive research using both observations and simulations has investigated secondary influences on galaxy distribution, such as assembly bias (Y.-T. Lin et al. 2016; R. H. Wechsler & J. L. Tinker 2018; A. Barreira, T. Lazeyras & F. Schmidt 2021; X. Xu, I. Zehavi & S. Contreras 2021a), inspiring multiple extensions to the baseline HOD formalism.

DESI employs a specific HOD model to generate its mock galaxy catalogues. In this work, we also explore variations in the HOD parameters to investigate how higher order clustering statistics can improve constraints on the galaxy–halo connection. Following the methodology used in the construction of the official DESI mocks, we use the highly efficient ABACUSHOD module to populate halo catalogues from the ABACUSSUMMIT simulations with mock galaxies. In addition to the baseline HOD prescription, the code supports a wide range of extensions, including galaxy assembly bias, velocity bias, and satellite profile flexibilities (S. Yuan et al. 2018; S. Yuan et al. 2022). The ABACUSHOD code is publicly distributed as part of the abacusutils package at <https://github.com/abacusorg/abacusutils>, with further documentation available at <https://abacusutils.readthedocs.io/en/latest/hod.html>.

4.2.1 Baseline model

For LRG tracers, the baseline HOD model follows Z. Zheng, A. L. Coil & I. Zehavi (2007):

$$\bar{n}_{\text{cent}}(M) = \frac{f_{\text{ic}}}{2} \left(1 + \text{erf} \left[\frac{\log_{10}(M/M_{\text{cut}})}{\sqrt{2}\sigma} \right] \right), \quad (5)$$

$$\bar{n}_{\text{sat}}(M) = \left[\frac{M - \kappa M_{\text{cut}}}{M_1} \right]^\alpha \bar{n}_{\text{cent}}(M), \quad (6)$$

with five baseline parameters M_{cut} , M_1 , σ , α , and κ . M_{cut} is the characteristic minimum mass of haloes that can host a central galaxy. σ is the width of the cut-off profile and controls the steepness of the

transition from 0 to 1 in the number of centrals. The number of satellite galaxies can be approximately described with a power law, where M_1 characterizes the typical halo mass that hosts one satellite galaxy and α is the power-law index. κ , together with M_{cut} , sets a mass threshold under which the probability of hosting a satellite is zero. We note that there is an extra term of $\bar{n}_{\text{cent}}(M)$ in the satellite occupation function to ensure that a halo can only host satellites if there exists a central galaxy in it (H. Guo et al. 2015b).

The DESI mocks have also included an incompleteness parameter f_{ic} , which controls the overall number density of the mock galaxies to match the observed galaxy number density. By definition, $0 \leq f_{\text{ic}} \leq 1$.

4.2.2 Extensions

Beyond the baseline model, ABACUSHOD supports multiple physically motivated HOD extensions, targeting diverse cosmological analyses. In this section, we briefly summarize several relevant extensions for the LRGs, which are employed in the construction of DESI mocks and may introduce variations in our analysis.

In addition to determining the expected number of galaxies per halo, the HOD framework also specifies the spatial and velocity distributions of galaxies within haloes. Because DESI models the full 2D redshift–space correlation function $\xi(r_p, \pi)$, which is sensitive to galaxy velocities, a velocity bias model is incorporated. Following H. Guo et al. (2015b), this model introduces two additional velocity bias parameters, α_c and α_s , to allow the central and satellite velocities to deviate from the velocity of underlying dark matter, offering more flexibility in modelling redshift–space clustering.

With this extension, the HOD model for LRGs is fully specified with eight parameters: the five baseline parameters M_{cut} , M_1 , σ , α , and κ ; the incompleteness parameter f_{ic} ; and the velocity bias parameters α_c and α_s . Redshift–space galaxy autocorrelation functions from the DESI One-Percent Survey (part of the Survey Validation from the Early Data Release) show that this model sufficiently describes LRG clustering, with no evidence for secondary biases or modifications from satellite profiles at the corresponding level of uncertainty (S. Yuan et al. 2024). As a result, DESI adopts this eight-parameter HOD model to construct the second-generation mocks, which we further describe in Section 4.3.

While the redshift–space 2PCF appears to be well described by the eight-parameter HOD model, we are interested in testing whether the more informative higher order statistic multiplet clustering, can further break degeneracies in HOD models that incorporate secondary biases. These biases introduce additional dependencies in the galaxy–halo connection beyond halo mass, potentially revealing more detailed aspects of galaxy formation physics. ABACUSHOD provides flexibility to include two forms of secondary bias: one based on halo concentration, and the other based on overdensity in the local environment. In the case of concentration-based bias, the HOD model incorporates the galaxy assembly history of the halo via the parameter of halo concentration (R. H. Wechsler et al. 2006; L. Gao & S. D. M. White 2007). Alternatively, environment-based secondary bias introduces a dependence on the dark matter density around each halo, allowing galaxy occupation to vary with the local environment (X. Xu, I. Zehavi & S. Contreras 2021b).

Following X. Xu et al. (2021b), ABACUSHOD models the biases by analytically mixing the secondary properties with the two mass parameters for centrals and satellites in the baseline model,

$$\log_{10} M_{\text{cut}}^{\text{mod}} = \log_{10} M_{\text{cut}} + A_c(c^{\text{rank}} - 0.5) + B_c(\delta^{\text{rank}} - 0.5), \quad (7)$$

¹For more details, see <https://abacussummit.readthedocs.io/en/latest/abacussummit.html>.

$$\log_{10} M_1^{\text{mod}} = \log_{10} M_1 + A_s(c^{\text{rank}} - 0.5) + B_s(\delta^{\text{rank}} - 0.5), \quad (8)$$

where c denotes the halo concentration parameter and δ represents the overdensity in the local dark matter distribution. Both properties are ranked in narrow halo mass bins and normalized to c^{rank} and δ^{rank} , which range from 0 to 1. Then, the four parameters (A_c , A_s , B_c , B_s) quantify the strength of the secondary bias in the HOD model. Physically, a positive A indicates a preference for lower concentration haloes, and a positive B indicates a preference for haloes in less dense environments. In the absence of secondary biases, all four parameters are zero.

4.3 DESI second-generation mocks

The DESI second-generation mocks are constructed from the ABACUSSUMMIT N -body simulations, specifically the 25 realizations at the Planck 2018 fiducial Λ CDM cosmology (Planck Collaboration VI 2020). The simulation outputs are organized into discrete snapshots at different redshifts, and for each tracer, HOD models calibrated to the measured clustering from the One-Percent Survey are applied to simulate proper galaxy samples from these underlying dark matter density fields. For LRGs, tracers at $z < 0.6$ are simulated using the periodic boxes at $z = 0.5$, while those at $z > 0.6$ are modelled with the $z = 0.8$ snapshot. In both cases, HOD parameters are calibrated using the eight-parameter model described in Section 4.2. DESI then takes the best-fitting HODs across different redshift snapshots and constructs ‘cutsy’ mocks by converting coordinates in periodic boxes to celestial coordinates and redshifts. Redshift–space distortions are incorporated by adjusting the redshifts of simulated galaxies accordingly.

DESI has developed and applied various treatments to account for survey systematics in the mock catalogues. In this work, we focus on two types of mocks: one without fibre assignment and one with it implemented. The *complete* mocks replicate the footprint of Y3 observations and the target selection masks, but no additional fibre assignment choices are incorporated. Thus, these mocks are treated as complete samples, and all completeness weights are 1. In contrast, the *altmtl* mocks are processed through the same fibre assignment pipeline applied for the actual observations, utilizing the real observational record and the same hardware status (J. Lasker et al. 2025). We expect these mocks to realistically reproduce survey systematics and enable unbiased clustering comparisons to observations. Similar to the observational data, the mocks are organized into LSS catalogues, with corresponding random catalogues generated via the LSS pipeline. The resulting catalogues closely match the actual data in footprint, redshift distribution, and number density, enabling consistent application of any clustering analysis pipeline to both the mocks and the observations.

The second-generation mocks used in this work are the same as those employed in the DESI DR1 key projects (DESI Collaboration 2025b, c), but with the updated Y3 footprint. More details on the simulations and the construction of these mocks are available in C. Zhao et al. (in preparation), and we will further discuss our mock choices in Section 5.

5 CALIBRATION FOR FIBRE ASSIGNMENT

Because DESI is a fibre-based instrument, fibre assignment procedures and fibre collisions result in unavoidable incompleteness in our sample. For example, in DESI, if there is a high-redshift ($z > 2.1$) quasar in the field, it will get priority during fibre assignment, and

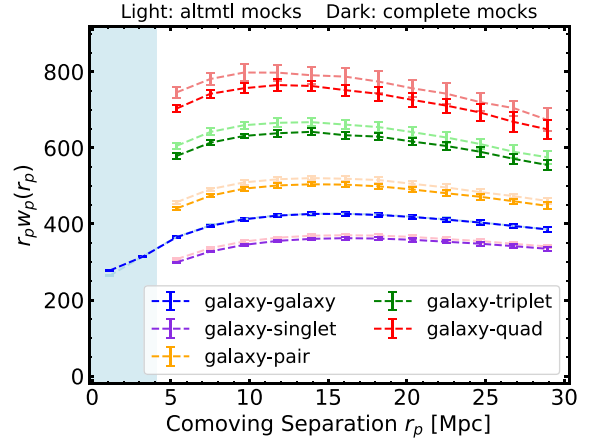


Figure 4. Comparison between the autocorrelations and the multiplet cross-correlations calculated from mocks without and with fibre assignment effects (the *complete* and *altmtl* mocks). The colour differences follow the convention of this paper, while dark colours mark results from the *complete* mocks and light colours represent results from the *altmtl* mocks.

thus nearby LRGs will not be observed (DESI Collaboration 2016a). Also, as in other fibre-fed surveys, the physical size and number density of fibres limit the survey to record redshifts of galaxies too close to each other on the focal plane (C. Hahn et al. 2017). DESI fibre collision effects are even more complicated as the fibres are all robotically controlled. These effects have a non-trivial impact on clustering measurements. Especially as we are measuring galaxy-multiplet cross-correlations, this incompleteness in observations can cause missing objects in identified galaxy multiplets (i.e. a quad appears as a triplet, a triplet appears as a pair) and thus lead to bias in the higher order statistics.

Therefore, we present a direct empirical comparison of how the observations would appear without and with fibre assignment effects, using the *complete* and *altmtl* mocks. Specifically, we compute the correlation functions using each of the 25 *complete* and *altmtl* mocks and use the averages of the 25 realizations as benchmarks for modelled clustering without or with fibre assignment implementation.

Fig. 4 compares the autocorrelations and the cross-correlations with singlets, pairs, triplets, and quads calculated from the *complete* and *altmtl* mocks, with error bars representing the standard deviation across the 25 realizations. While the *altmtl* mocks exhibit a lower galaxy number density (approximately 88%) due to incomplete observations, the redshift distributions remain consistent between the *complete* and *altmtl* mocks, and the autocorrelation functions show good agreement on large scales. This consistency suggests that DESI’s weighting scheme effectively calibrates survey systematics in autocorrelation measurements as expected.

Beyond the autocorrelations, the multiplet cross-correlations are expected to be more sensitive to survey geometry. Fig. 4 shows that the multiplet clustering amplitude inferred from the *complete* mocks is consistently lower than that measured with the *altmtl* mocks, indicating that when there is no additional assembly bias, a complete sample favours less massive dark matter haloes. This makes sense because, in the *altmtl* mocks, fibre assignment can result in missing galaxies from a multiplet, causing the remaining system to exhibit larger clustering bias than multiplets of the same size in the *complete* mocks. For example, a triplet from the *altmtl*

mocks can potentially be a quad, which correlates preferentially with a more massive halo. Similarly, while a singlet in the complete mocks truly represents an isolated galaxy, a singlet in the *altmtl* mocks may arise from a higher multiplicity system with unobserved members.

Because incomplete observations can artificially enhance the amplitudes of galaxy-multiplet cross-correlations, calibrating the impact of fibre assignment on multiplet clustering is necessary. As generating proper light-cones and cutsky mocks that match survey specifics is complicated and computationally expensive, this comparison between complete and *altmtl* mocks also serves as a nice empirical calibration for fibre assignment effects for later analysis in Section 6.2. We use the difference in clustering amplitudes, averaged over large scales ($r_p > 10$ Mpc), to correct for survey systematics before comparing observations with mocks generated from periodic boxes. While we understand this calibration is not perfect, as the second-generation mocks introduce their own systematics, it provides a simple benchmark to understand and isolate the effect of survey specifics, effectively eliminating contamination from survey geometry to the physics of galaxy–halo connections.

6 THE GALAXY–DARK MATTER CONNECTION

To understand the physics of the galaxy–dark matter connection, we need to compare clustering statistics from observation and simulation. In this section, we present comparisons of multiplet clustering measured from the observations with those calculated from the DESI second-generation mocks and from mocks with modified halo occupation models.

6.1 Comparison between DESI mocks and observations

The DESI second-generation mocks are constructed using the HOD model with the best-fitting parameters for the 2PCFs from the One-Percent Survey, while the *altmtl* mocks have been processed through the fibre assignment pipeline that closely replicates actual survey procedures. Therefore, these mocks offer robust theoretical predictions for galaxy clustering and are invaluable for interpreting DESI observations. Accordingly, we first compare the observed clustering with modelled predictions from the *altmtl* mocks, as they are supposed to incorporate both galaxy assembly physics and realistic survey systematics.

We identify galaxy multiplets using galaxy catalogues from each of the 25 *altmtl* mocks and compute both the autocorrelation functions and the multiplet cross-correlation functions, following the methodology described in Section 3. As these mocks are organized into LSS catalogues with corresponding randoms, all pair counts are weighted in the same way as for galaxies from observations. For all clustering measurements, we use the average of the 25 *altmtl* mocks as the final prediction and compute the error bars as the standard deviation across the 25 realizations in each r_p bin.

Fig. 5 compares the multiplet cross-correlations calculated from observations and the *altmtl* mocks, including autocorrelations as a benchmark. Observational results are shown in darker solid lines, while those from the mocks are plotted in lighter colours. We note that the error bars for the mocks differ from those of the observations: for the latter, we estimate uncertainties using jackknife resampling, while for the mocks, we use the scatter among independent realizations. The plot shows that the autocorrelation from the mocks agrees reasonably well with the observations, indicating that the HOD model used in the second-generation mocks provides a good description of the projected correlation function and is reasonable to use in DESI

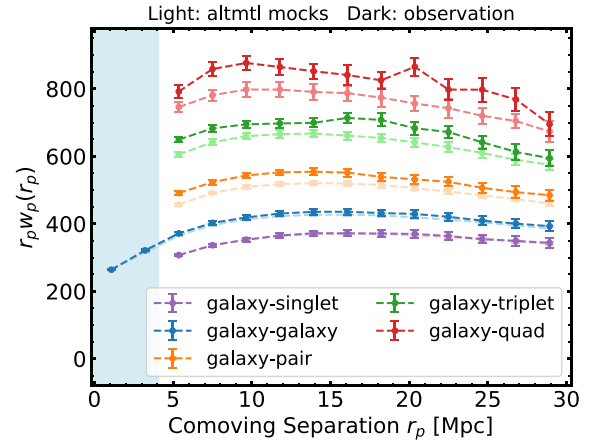


Figure 5. Comparison between the galaxy-multiplet cross-correlations measured from the Y3 observations and the *altmtl* mocks. Following Fig. 3, each colour represents clustering measured with galaxy multiplets of different sizes, with lighter colours indicating the mock results and darker colours depicting the observations.

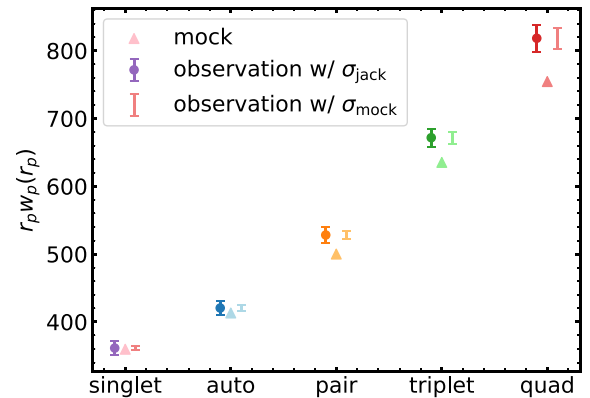


Figure 6. Comparison between the general clustering amplitudes for galaxies and multiplets, represented by the projected correlation functions averaged over a specific r_p range. The triangular points in lighter colours show results from the mocks, while darker circular points depict the observations. Two versions of error bars are included: jackknife errors estimated from the observations are shown on the left, and error bars capturing the scatter between mocks are shown on the right.

cosmological analyses. However, while the HOD model successfully reproduces a 2PCF relatively consistent with observations, this agreement does not extend to the cross-correlations with galaxy multiplets, where significant discrepancies appear. In particular, the galaxy multiplets in the mocks exhibit smaller clustering bias than those in the actual observations. If there is no additional assembly bias, this difference in the clustering amplitudes would suggest that mock multiplets tend to trace less massive dark matter haloes; otherwise, there could be additional assembly bias specific to multiplets that alters clustering.

To condense the information from the projected correlation functions and facilitate the comparison between observations and mocks, we average $r_p w_p$ over the scale $10 \text{ Mpc} < r_p < 30 \text{ Mpc}$ to obtain the general clustering amplitudes. Fig. 6 presents these condensed clustering statistics for both observations and mocks, including the autocorrelations as well as the cross-correlations with galaxy singlets, pairs, triplets, and quads. Here, we include two versions of error

bars on the observations, one from jackknife resampling of the data and the other purely derived from the mocks. Specifically, σ_{jack} is calculated with the standard jackknife procedure (Equation 4) based on the average clustering amplitudes in the 50 jackknife replicates described in Section 3, while σ_{mock} is obtained as the standard deviation of the 25 average clustering amplitudes measured from the 25 mocks. We note that the jackknife error bars estimated with data consistently exceed the scatter between mocks, particularly for the cross-correlation with singlets and the galaxy autocorrelation. We repeat the same jackknife procedure on the mocks and find error bars of similar amplitudes to those from the real observations, suggesting that the larger variance is not due to unknown intrinsic fluctuations in the data but is instead a feature of the jackknife method itself. At the same time, the smaller scatter among the mocks suggests that the mock catalogues are not subject to additional sources of systematic uncertainty. Understanding that the jackknife method can be conservative, we include both the jackknife and the mock-based error bars to provide a more complete characterization of uncertainties in our measurements.

Despite the differences between the two error estimates, we observe statistically significant discrepancies at a few-sigma level between the mocks and observations for the multiplet cross-correlations, with multiplets in the simulations exhibiting smaller clustering amplitudes. The HOD fails to predict the higher order correlations as effectively as it does for the autocorrelation. These discrepancies suggest that the vanilla HOD model is insufficient to describe the behaviour of close galaxy multiplets, highlighting the potential of multiplet clustering as a higher order statistic to offer better constraints on the galaxy–halo connection.

We also directly compare the number densities of multiplets in the observations and mocks to complement the multiplet clustering analysis. The fraction of pairs among all galaxies in the mocks is about 5 per cent higher than in observations, the number density of triplets is about 1 per cent higher, and the observed number density of quads is consistent with statistical fluctuations in the mock predictions (about 1 per cent). These discrepancies suggest that the mocks are slightly off-tuned, with the HOD performing less well for lower mass haloes. However, compared to the disagreement in clustering amplitudes, the differences in number densities are less significant and physically different, highlighting the distinct physics multiplet clustering probes and motivating additional assembly bias for the multiplets.

6.2 Insights on secondary bias

There have been studies showing that standard HOD models applied to cosmological N -body simulations may not be able to fully describe the galaxy distribution in hydrodynamical simulations that incorporate more complex physics of galaxy formation (i.e. B. Hadzhiyska et al. (2020)). Nevertheless, employing N -body simulations and modelling the galaxy–halo connection with HOD is computationally efficient, and thus it is worth testing whether additional flexibility in the HOD can improve the performance of this simple but powerful approach. At the same time, the various discrepancies between observations and mocks in different multiplet clustering measurements are reassuring that these higher order cross-correlations can provide more information about how galaxies trace dark matter.

As an efficient higher order statistic, multiplet clustering offers valuable opportunities to study the effect of secondary halo properties on the galaxy–halo connection. While in this paper, we will not attempt to constrain the HOD parameters through full statistical

inferences, this section demonstrates how this novel statistic can break degeneracies in galaxy–halo connection models and provide insights into secondary biases through testing the reconstruction of multiplet cross-correlations under various modifications to the vanilla HOD model.

We utilize the cubic box at Planck 2018 cosmology from the ABACUSSUMMIT simulation suite, specifically the redshift snapshot at $z = 0.5$, as the model template for the underlying dark matter density field. We then generate mock galaxy catalogues with different HOD prescriptions using the ABACUSHOD package and incorporate redshift–space distortions along the z -axis. The incompleteness parameter, f_{ic} , is tuned to match the number density with that of the complete mocks. To consistently identify galaxy pairs and multiplets in the periodic box as in real observations, we use the z -axis as the LOS and apply periodic wrapping conditions to ensure that galaxies interact with each other in a continuous system. We set the separation threshold to 10 Mpc in 3D and 1.5 Mpc in the x – y plane for the selection of galaxy pairs, which are linked into multiplets using the friends-of-friends algorithm. These multiplets are then used to compute the cross-correlation functions, and we calibrate the average clustering amplitudes following Section 5 before comparing with real observations. To be consistent with the simulation snapshot at $z=0.5$, the observational measurements and mock-based calibrations in this section use only galaxies in the redshift range of $0.4 < z < 0.6$.

Although the projected correlation function w_p is largely insensitive to velocity bias or satellite profile parameters, these factors could potentially affect multiplet identification due to redshift space distortions. Therefore, before exploring assembly bias, we test whether satellite velocity and spatial biases can alter multiplet clustering signals and improve the agreement with observations.

The second-generation mocks are generated with the inclusion of satellite velocity bias, carefully calibrated to match the full-shape correlation function $\xi(r_p, \pi)$ down to small scales. Thus, the fact that the observed clustering of multiplets differs from the mocks already indicates that the velocity bias alone might be insufficient to explain the discrepancy identified in Section 6.1. To further support this conclusion and explore the potential impact of satellite velocity bias, we fix the five baseline HOD parameters and vary the satellite velocity bias parameter α_s across a wide range from 0.5 to 1.2. While this range spans physically very distinct regimes, well beyond the typical values measured from DESI, similar samples, and hydrodynamical simulations (H. Guo et al. 2015a; J.-N. Ye et al. 2017), it induces only mild variations in multiplet clustering amplitudes, with all differences remaining below the 1σ level. Moreover, we find that the number density of multiplets is more sensitive to α_s than the clustering amplitudes are, suggesting that satellite velocity bias can be well constrained through careful calibrations of number densities in addition to small-scale full-shape clustering. Thus, the influence of satellite velocity bias is effectively limited for our analysis.

The case for satellite spatial bias is more straightforward. Varying the satellite spatial bias parameter s from -0.3 to 0.3 results in only minor variations in the amplitudes of multiplet clustering, allowing us to safely exclude satellite spatial bias from further consideration.

We remind that these tests do not necessarily imply that multiplet clustering lacks constraining power for velocity bias or satellite profiles. In fact, we expect that 2D full-shape cross-correlation functions with multiplets can encode valuable information about these parameters. However, such analyses are beyond the scope of this work, as we are presenting a highly compressed statistic, the average clustering amplitude from the projected correlation function, to provide simple demonstrations of potential future applications.

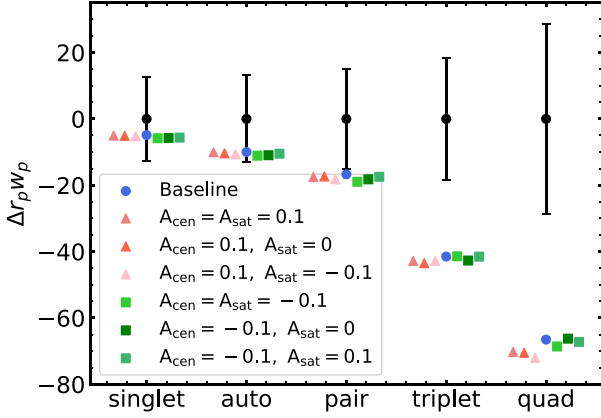


Figure 7. The average clustering amplitudes measured from observations and predicted by HOD models with different concentration-based secondary biases added to the same set of baseline parameters. Observational data points with jackknife error bars are fixed at zero as references. Models with different secondary biases, as indicated in the labels, are plotted from left to right on either side of the baseline model results.

Since assembly bias is important for the galaxy–halo connection and early DESI data have limited statistical power to distinguish between models with different assembly biases, we choose to demonstrate the impact of multiplet clustering specifically on the secondary biases.

We focus on two physically motivated HOD extensions that capture the impact of halo assembly history and local environment on the galaxy–halo connection, following the implementation described in Section 4.2. To ensure consistency, we first tune a vanilla HOD model to match the 2PCF of the complete mocks, and refer to this model as the ‘baseline’ throughout the remainder of this section.

First, we experiment with the concentration-based secondary bias described by A_c and A_s . Fig. 7 showcases the deviation in the two-point autocorrelations and the multiplet cross-correlations with the inclusion of different (A_c, A_s) on top of the baseline parameters. We compare the differences in clustering amplitudes ($\Delta r_p w_p$) between the models, including the baseline and those with concentration-based biases, and the observations, which are centred at zero with jackknife error bars indicating uncertainty. The plot demonstrates that the inclusion of (A_c, A_s) on top of the same set of baseline HOD parameters does not significantly affect the autocorrelation, suggesting that the two-point function alone is not sensitive enough to distinguish between models with different concentration-based biases. The impact of (A_c, A_s) on higher order cross-correlations is larger. However, compared with the increasing error bars for correlations with larger multiplets and the discrepancy between mocks and observations, the improvement in the differentiating power of multiplet clustering does not seem statistically important.

We further increase the magnitudes of the secondary bias parameters A_c and A_s to induce a larger deviation from the autocorrelation predicted by the baseline model, thereby enhancing the dependence on halo concentration and ensuring their statistical significance. We then perform a systematic grid search over the five vanilla HOD parameters around their baseline values at two sets of fixed (A_c, A_s) , selecting the combination of parameters that best fits the baseline two-point function based on the χ^2 calculated using the jackknife error bars from observations. The resulting shifts in vanilla HOD parameters remain small after introducing a more prominent dependence on halo concentration, and Fig. 8 illustrates the modifications of multiplet clustering with the inclusion of these

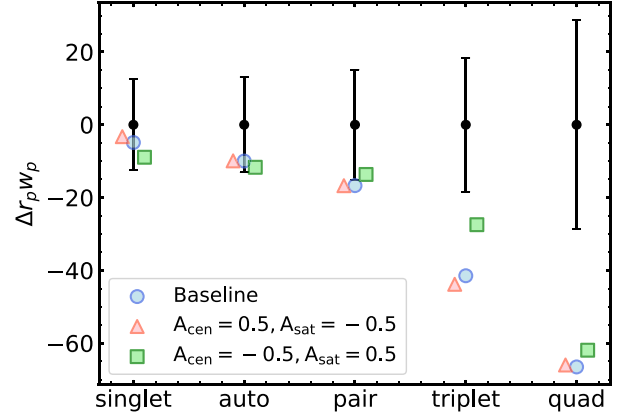


Figure 8. Example of average clustering amplitudes measured from observations and predicted by HOD models incorporating different concentration-based secondary biases of enhanced magnitudes. Observational data points with jackknife error bars are fixed at zero as references. Predictions from the baseline HOD model and from models incorporating positive and negative A_c are represented by circles, triangles, and squares, respectively.

stronger concentration-based biases. While the agreement between the model and observations can be improved, the overall conclusion remains that the impact of concentration-based secondary bias is not statistically significant enough and including (A_c, A_s) does not resolve the discrepancies observed in higher order clustering.

Another set of secondary biases we explore is the environment-based bias described by B_c and B_s . Again, we start by including different values of (B_c, B_s) on top of the same set of baseline HOD parameters. However, this time we observe large deviations in the predictions of autocorrelations from different models, which is consistent with previous work that vanilla HOD parameters are more sensitive to the inclusion of environment-based bias. To recover a reasonable fit to the two-point function so that the comparison of multiplet clustering predictions is meaningful, we again perform a systematic grid search around the baseline HOD parameters at fixed values of (B_c, B_s) . We explore two simplified scenarios, $B_c = B_s = \pm 0.05$, and terminate the grid search once the minimum χ^2 calculated with jackknife error bars falls below a certain threshold. Following this procedure, we use the best parameters from the grid search to generate two HOD models with different environment-based secondary biases, while both models predict reasonable galaxy–galaxy autocorrelations. Moreover, as we modify the bias parameters systematically, we can infer the relation between HOD parameters. For example, when including negative (B_c, B_s) , we see an increase in σ and a decrease in M_1 compared to the original baseline parameters, which translates to moving galaxies to lower mass haloes in denser environments and is consistent with previous work based on complete statistical inference (S. Yuan et al. 2022).

Given that these models can fit the observed autocorrelation equally well, we treat them as benchmarks and evaluate their performance using higher order statistics. Fig. 9 compares the averaged autocorrelations and cross-correlations from observations and mocks incorporating different environment-based secondary biases. While all three models provide reasonable fits to the autocorrelation, their performance diverges when describing the galaxy–multiplet cross-correlations. We find that these deviations in multiplet clustering predictions from different environment-based biases are much more prominent than those from different concentration-based biases. Moreover, including positive or negative environment-based biases

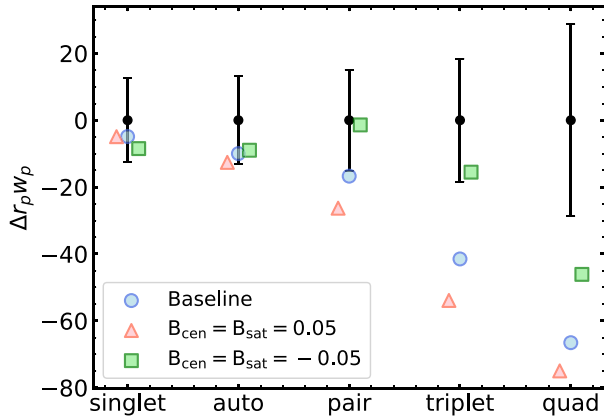


Figure 9. Similar to Fig. 8 but for different environment-based secondary biases (B_c , B_s).

shifts the amplitudes of the cross-correlations in opposite directions. Notably, the model with negative (B_c , B_s) significantly improves the consistency between mocks and observations for the multiplet cross-correlations. Therefore, we suggest that environment-based secondary bias is potentially important for fully capturing the physics of the galaxy–halo connection and that multiplet clustering offers a unique opportunity to reveal this preference.

It is worth noting that the results presented in this section are primarily methodology-oriented. Our aim is not to constrain the HOD parameters directly, but rather to provide insights into the use of higher order statistics and the inclusion of secondary biases when modelling galaxy clustering. We demonstrate that incorporating galaxy assembly bias into the HOD model, particularly environment-based secondary biases, has the potential to provide a more accurate description of the measured galaxy–multiplet clustering. Although definitive physical conclusions would require a full statistical inference on the data, we observe a trend that LRGs favourably occupy haloes in denser environments, as indicated by the preference for negative environment-based secondary bias. Furthermore, we highlight the potential of utilizing multiplet clustering to enhance the constraints on galaxy–dark matter connections. When models produce similar predictions for the two-point autocorrelation, making them difficult to distinguish due to statistical uncertainties, incorporating galaxy–multiplet cross-correlations in the analysis is helpful to break model degeneracies. As DESI continues to survey larger sky areas with higher completeness, the errors in this analysis are expected to decrease significantly, further strengthening the constraining power of multiplet clustering.

7 DISCUSSIONS

In this work, we propose multiplet clustering as an efficient higher order statistic for analysing galaxy clustering. Using LRGs from DESI DR2, we identify small groups of nearby galaxies, referred to as multiplets, and cross-correlate the resulting multiplet field with the general galaxy field to measure galaxy–multiplet clustering. While this estimator is computationally much cheaper than conventional higher order statistics such as the 3PCF or bispectrum, it still captures higher order information and small-scale non-linearities, offering sensitivity to galaxy bias and cosmology. Specifically, we cross-correlate galaxy singlets, pairs, triplets, and quads with the general galaxy field to obtain the galaxy–multiplet cross-correlation functions. We observe larger clustering bias for larger multiplets as

tracers of more massive dark matter haloes, suggesting that these multiplets trace the LSS differently than individual galaxies.

We compare the observed clustering statistics with mock catalogues constructed using the cosmological N -body simulation ABACUS-SUMMIT with HOD models describing the galaxy–halo connection. While the autocorrelation measured with the official DESI second-generation mocks agrees well with actual observations, the multiplet cross-correlations from the data exhibit higher amplitudes compared with those predicted by the mocks (highlighted in Fig. 5 and Fig. 6). This discrepancy indicates that, following the simplest assumptions, the multiplets in the mocks preferentially occupy less massive haloes than those from observations, or that there could be additional assembly bias specific to the multiplets, suggesting the importance of higher order statistics in capturing the full complexity of the galaxy–halo connection and motivating the exploration of extended HOD models that include secondary biases. We test two physically motivated HOD extensions dependent on secondary halo properties: halo concentration and local environment. While these models are calibrated to match the same 2PCF, they produce different predictions for the galaxy–multiplet cross-correlations. We showcase that including secondary biases can potentially describe the higher order correlations better, and the HOD model with environment-based secondary bias generates the most significant improvement in the data–model agreement, indicating that galaxies prefer to occupy haloes in denser environments.

We highlight the potential of using multiplet clustering to break degeneracies in models describing galaxy bias. Even though we have not performed a full statistical inference on the data, we observe that different models with the same descriptive power for the autocorrelation diverge in their predictions for the galaxy–multiplet cross-correlations. Therefore, the employment of these higher order correlation functions, beyond providing information on how galaxy multiplets of various sizes trace dark matter differently, offers valuable opportunities to distinguish between complex models of the galaxy–dark matter connection.

While we establish a good foundation for multiplet clustering analysis, the physical conclusions for the galaxy–halo connection are not yet comprehensive because the HOD modelling presented in this work does not utilize a quantitative likelihood analysis, which is essential for comparing HOD models and constraining the parameters statistically. Based on simple observations of various performances of HODs with different model extensions, we anticipate that a complete statistical inference of HOD models utilizing the multiplet clustering statistics, along with careful covariance matrix calculations and a larger data sample from future DESI observations, will provide powerful insights into galaxy assembly bias and potentially cosmology. Furthermore, we plan to investigate the full-shape correlation as it offers more information in the redshift space. We also hope to extend the analysis to other classes of galaxies beyond LRGs, considering their distinct physical properties.

Beyond clustering analysis, the multiplets themselves encode interesting physics and offer opportunities for many other cosmological studies. For example, we can investigate how halo mass correlates with multiplet properties, such as internal separations in real space and redshift space, to further understand how galaxy multiplets of various properties trace the LSS differently. Moreover, because isolated galaxies and galaxies residing in multiplets are tracers of the same underlying matter distribution but have different biases, they are potentially useful for multitracers cosmology analysis (P. McDonald & U. Seljak 2009), where sample variance can be

effectively suppressed, allowing for more precise measurements of redshift–space distortions.

We can also stack weak lensing signals around the multiplets and compare them to galaxy–galaxy lensing measurements to probe the matter distribution of the universe. While joint analyses between galaxy clustering and galaxy–galaxy lensing have proven valuable for reconstructing galaxy bias (A. Leauthaud et al. 2012; S. More et al. 2013a, b), extending such analyses to another class of objects, such as galaxy multiplets, can further enhance our understanding of the galaxy–dark matter connection and improve constraints on cosmological parameters such as the amplitude of matter clustering σ_8 . Specifically, adding lensing information allows us to calibrate the matter distribution around multiplets and verify the source of their enhanced clustering signals. Moreover, an observational tension exists between galaxy clustering and galaxy–galaxy lensing (A. Leauthaud et al. 2017), and S. Yuan et al. (2021) argue that incorporating galaxy assembly bias into the HOD framework can help explain this lensing tension. With extra information provided by the multiplet statistics, we can further evaluate this tension and explore possible solutions.

Finally, we would like to explore cross-correlations with surveys of the Sunyaev–Zeldovich effect as another probe of the density distribution of the universe while investigating the thermal and dynamical properties of the multiplets.

ACKNOWLEDGEMENTS

We thank the anonymous referee for the valuable comments and suggestions. We thank Johannes Lange and Ying Zu for their insightful feedback. HW was supported by the Herchel Smith Undergraduate Science Research Program at Harvard College and the Stanford Graduate Fellowship. DJE was supported by U.S. Department of Energy grant DE-SC0007881 and as a Simons Foundation Investigator. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231.

This material is based upon work supported by the U.S. Department of Energy (DOE), Office of Science, Office of High-Energy Physics, under Contract No. DE-AC0205CH11231, and by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract. Additional support for DESI was provided by the U.S. National Science Foundation (NSF), Division of Astronomical Sciences under Contract No. AST-0950945 to the NSF’s National Optical-Infrared Astronomy Research Laboratory; the Science and Technology Facilities Council of the United Kingdom; the Gordon and Betty Moore Foundation; the Heising–Simons Foundation; the French Alternative Energies and Atomic Energy Commission (CEA); the National Council of Humanities, Science and Technology of Mexico (CONAHCYT); the Ministry of Science, Innovation and Universities of Spain (MICIU/AEI/10.13039/501100011033), and by the DESI Member Institutions: <https://www.desi.lbl.gov/collaborating-institutions>. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U. S. National Science Foundation, the U. S. Department of Energy, or any of the listed funding agencies.

The authors are honored to be permitted to conduct scientific research on I’oligam Du’ag (Kitt Peak), a mountain with particular significance to the Tohono O’odham Nation.

DATA AVAILABILITY

Loa data will be released as part of the DESI Data Release 2 (DR2). Data products from the ABACUSSUMMIT simulation are available at <https://abacusnbody.org>. All data points plotted in this paper are available in Zenodo at <https://zenodo.org/records/17648090>.

REFERENCES

- Anand A. et al., 2024, *AJ*, 168, 124
 Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, *ApJ*, 304, 15
 Barreira A., Lazeyras T., Schmidt F., 2021, *J. Cosmol. Astropart. Phys.*, 2021, 029
 Bernardeau F., Colombi S., Gaztañaga E., Scoccimarro R., 2002, *Phys. Rep.*, 367, 1
 Cooray A., Sheth R., 2002, *Phys. Rep.*, 372, 1
 DESI Collaboration, 2016a, preprint (arXiv:1611.00036)
 DESI Collaboration, 2016b, preprint (arXiv:1611.00037)
 DESI Collaboration, 2022, *AJ*, 164, 207
 DESI Collaboration, 2024a, *AJ*, 167, 62
 DESI Collaboration, 2024b, *AJ*, 168, 58
 DESI Collaboration, 2024c, *J. Cosmol. Astropart. Phys.*, 2028, 008
 DESI Collaboration, 2025a, *Phys. Rev. D*, 112, 083515
 DESI Collaboration, 2025b, *J. Cosmol. Astropart. Phys.*, 2025, 012
 DESI Collaboration, 2025c, *J. Cosmol. Astropart. Phys.*, 2025, 017
 DESI Collaboration, 2025d, *J. Cosmol. Astropart. Phys.*, 2025, 028
 DESI Collaboration, 2025e, preprint (arXiv:2503.14745)
 Davis M., Peebles P. J. E., 1983, *ApJ*, 267, 465
 Dey A. et al., 2019, *AJ*, 157, 168
 Gao L., White S. D. M., 2007, *MNRAS*, 377, L5
 Garrison L. H., Eisenstein D. J., Ferrer D., Maksimova N. A., Pinto P. A., 2021, *MNRAS*, 508, 575
 Gaztañaga E., Norberg P., Baugh C. M., Croton D. J., 2005, *MNRAS*, 364, 620
 Guo H. et al., 2015a, *MNRAS*, 446, 578
 Guo H. et al., 2015b, *MNRAS*, 453, 4368
 Guy J. et al., 2023, *AJ*, 165, 144
 Hadzhiyska B., Bose S., Eisenstein D., Hernquist L., Spergel D. N., 2020, *MNRAS*, 493, 5506
 Hahn C., Scoccimarro R., Blanton M. R., Tinker J. L., Rodríguez-Torres S. A., 2017, *MNRAS*, 467, 1940
 Kulkarni G. V., Nichol R. C., Sheth R. K., Seo H.-J., Eisenstein D. J., Gray A., 2007, *MNRAS*, 378, 1196
 Lamman C. et al., 2024, *MNRAS*, 534, 3540
 Landy S. D., Szalay A. S., 1993, *ApJ*, 412, 64
 Lasker J. et al., 2025, *J. Cosmol. Astropart. Phys.*, 2025, 127
 Leauthaud A. et al., 2012, *ApJ*, 744, 159
 Leauthaud A. et al., 2017, *MNRAS*, 467, 3024
 Lin Y.-T., Mandelbaum R., Huang Y.-H., Huang H.-J., Dalal N., Diemer B., Jian H.-Y., Kravtsov A., 2016, *ApJ*, 819, 119
 Maksimova N. A., Garrison L. H., Eisenstein D. J., Hadzhiyska B., Bose S., Satterthwaite T. P., 2021, *MNRAS*, 508, 4017
 Marín F., 2011, *ApJ*, 737, 97
 McBride C. K., Connolly A. J., Gardner J. P., Scranton R., Newman J. A., Scoccimarro R., Zehavi I., Schneider D. P., 2011, *ApJ*, 726, 13
 McDonald P., Seljak U., 2009, *J. Cosmol. Astropart. Phys.*, 2009, 007
 Miller T. N. et al., 2024, *AJ*, 168, 95
 More S., van den Bosch F. C., Cacciato M., More A., Mo H., Yang X., 2013a, *MNRAS*, 430, 747
 More S., van den Bosch F. C., Cacciato M., More A., Mo H., Yang X., 2013b, *MNRAS*, 430, 747
 Myers A. D. et al., 2023, *AJ*, 165, 50
 Peebles P. J. E., 1980, The Large-scale Structure of the Universe.
 Planck Collaboration VI, 2020, *A&A*, 641, A6
 Poppett C. et al., 2024, *AJ*, 168, 245
 Ross A. J. et al., 2025, *J. Cosmol. Astropart. Phys.*, 2025, 125
 Schlafly E. F. et al., 2023, *AJ*, 166, 259

- Silber J. H. et al., 2023, *AJ*, 165, 9
- Sinha M., Garrison L. H., 2020, *MNRAS*, 491, 3022
- Slepian Z. et al., 2017, *MNRAS*, 469, 1738
- Wechsler R. H., Tinker J. L., 2018, *ARA&A*, 56, 435
- Wechsler R. H., Zentner A. R., Bullock J. S., Kravtsov A. V., Allgood B., 2006, *ApJ*, 652, 71
- Wright E. L. et al., 2010, *AJ*, 140, 1868
- Xu X., Zehavi I., Contreras S., 2021a, *MNRAS*, 502, 3242
- Xu X., Zehavi I., Contreras S., 2021b, *MNRAS*, 502, 3242
- Ye J.-N., Guo H., Zheng Z., Zehavi I., 2017, *ApJ*, 841, 45
- Yuan S., Eisenstein D. J., Garrison L. H., 2017, *MNRAS*, 472, 577
- Yuan S., Eisenstein D. J., Garrison L. H., 2018, *MNRAS*, 478, 2019
- Yuan S., Hadzhiyska B., Bose S., Eisenstein D. J., Guo H., 2021, *MNRAS*, 502, 3582
- Yuan S., Garrison L. H., Hadzhiyska B., Bose S., Eisenstein D. J., 2022, *MNRAS*, 510, 3301
- Yuan S. et al., 2024, *MNRAS*, 530, 947
- Zhang H. et al., 2022, *MNRAS*, 515, 6133
- Zheng Z. et al., 2005, *ApJ*, 633, 791
- Zheng Z., Coil A. L., Zehavi I., 2007, *ApJ*, 667, 760
- Zhou R. et al., 2023, *AJ*, 165, 58
- Zou H. et al., 2017, *PASP*, 129, 064101
- ¹Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA
- ²Department of Physics, Stanford University, 382 Via Pueblo Mall, Stanford, CA 94305, USA
- ³Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, 452 Lomita Mall, Stanford, CA 94305, USA
- ⁴Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA
- ⁵Department of Physics, Boston University, 590 Commonwealth Avenue, Boston, MA 02215, USA
- ⁶Dipartimento di Fisica ‘Aldo Pontremoli’, Università degli Studi di Milano, Via Celoria 16, I-20133 Milano, Italy
- ⁷INAF – Osservatorio Astronomico di Brera, Via Brera 28, I-20122 Milano, Italy
- ⁸Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK
- ⁹Instituto de Física, Universidad Nacional Autónoma de México, Circuito de la Investigación Científica, Ciudad Universitaria, Cd. de México C. P. 04510, México
- ¹⁰NSF NOIRLab, 950 N. Cherry Ave., Tucson, AZ 85719, USA
- ¹¹Department of Astronomy and Astrophysics, University of Toronto, Toronto, ON M5S 3H4, Canada
- ¹²Department of Physics and Astronomy and Pittsburgh Particle Physics, Astrophysics, and Cosmology Center (PITT PACC), University of Pittsburgh, 3941 O’Hara Street, Pittsburgh, PA 15260, USA
- ¹³University of California, Berkeley, 110 Sproul Hall #5800 Berkeley, CA 94720, USA
- ¹⁴Institut de Física d’Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Edifici Cn, Campus UAB, E-08193, Bellaterra (Barcelona), Spain
- ¹⁵Departamento de Física, Universidad de los Andes, Cra. 1 No. 18A-10, Edificio Ip, CP Bogotá 111711, Colombia
- ¹⁶Osservatorio Astronomico, Universidad de los Andes, Cra. 1 No. 18A-10, Edificio H, CP Bogotá 111711, Colombia
- ¹⁷Institut d’Estudis Espacials de Catalunya (IEEC), c/ Esteve Terradas 1, Edifici RDIT, Campus PMT-UPC, E-08860 Castelldefels, Spain
- ¹⁸Institute of Cosmology and Gravitation, University of Portsmouth, Dennis Sciama Building, Portsmouth PO1 3FX, UK
- ¹⁹Institute of Space Sciences, ICE-CSIC, Campus UAB, Carrer de Can Magrans s/n, E-08913 Bellaterra, Barcelona, Spain
- ²⁰Fermi National Accelerator Laboratory, PO Box 500, Batavia, IL 60510, USA
- ²¹Center for Cosmology and AstroParticle Physics, The Ohio State University, 191 West Woodruff Avenue, Columbus, OH 43210, USA
- ²²Department of Physics, The Ohio State University, 191 West Woodruff Avenue, Columbus, OH 43210, USA
- ²³The Ohio State University, Columbus, OH 43210, USA
- ²⁴Department of Physics, The University of Texas at Dallas, 800 W. Campbell Road, Richardson, TX 75080, USA
- ²⁵Department of Physics and Astronomy, University of California, Irvine, CA 92697, USA
- ²⁶Departament de Física, Serra Hùnter, Universitat Autònoma de Barcelona, E-08193 Bellaterra (Barcelona), Spain
- ²⁷Institució Catalana de Recerca i Estudis Avançats, Passeig de Lluís Companys, 23, E-08010 Barcelona, Spain
- ²⁸Department of Physics and Astronomy, University of Sussex, Brighton BN1 9QH, UK
- ²⁹Departamento de Física, DCI-Campus León, Universidad de Guanajuato, Loma del Bosque 103, León, Guanajuato C. P. 37150, México
- ³⁰Instituto Avanzado de Cosmología A. C., San Marcos 11 – Atenas 202, Magdalena Contreras. Ciudad de México C. P. 10720, México
- ³¹IRFU, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France
- ³²Department of Physics and Astronomy, University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada
- ³³Perimeter Institute for Theoretical Physics, 31 Caroline St North, Waterloo, ON N2L 2Y5, Canada
- ³⁴Waterloo Centre for Astrophysics, University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada
- ³⁵Instituto de Astrofísica de Andalucía (CSIC), Glorieta de la Astronomía, s/n, E-18008 Granada, Spain
- ³⁶Departament de Física, EEBE, Universitat Politècnica de Catalunya, c/Eduard Maristany 10, E-08930 Barcelona, Spain
- ³⁷Department of Astronomy, The Ohio State University, 4055 McPherson Laboratory, 140 W 18th Avenue, Columbus, OH 43210, USA
- ³⁸Department of Physics and Astronomy, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Republic of Korea
- ³⁹CIEMAT, Avenida Complutense 40, E-28040 Madrid, Spain
- ⁴⁰Department of Physics, University of Michigan, 450 Church Street, Ann Arbor, MI 48109, USA
- ⁴¹University of Michigan, 500 S. State Street, Ann Arbor, MI 48109, USA
- ⁴²National Astronomical Observatories, Chinese Academy of Sciences, A20 Datun Road, Chaoyang District, Beijing 100101, P. R. China

This paper has been typeset from a $\text{\LaTeX}$ file prepared by the author.