Detection of spatial clustering in the 1000 richest SDSS DR8 redMaPPer clusters with Nearest Neighbor distributions

Yunchong Wang, Arka Banerjee, and Tom Abel

1 Department of Physics, Indian Institute of Science Education and Research, Homi Bhabha Road, Pashan, Pune 411008, India
2 Kavli Institute for Particle Astrophysics & Cosmology, Stanford University, 452 Lomita Mall, Stanford, CA 94305, USA
3 SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA
4 Fermi National Accelerator Laboratory, Cosmic Physics Center, Batavia, IL 60510, USA
5 Department of Physics, Indian Institute of Science Education and Research, Homi Bhabha Road, Pashan, Pune 411008, India

Accepted ***. Received ***; in original form ***

ABSTRACT

Distances to the $k$-nearest-neighbor ($k$NN) data points from volume-filling query points are a sensitive probe of spatial clustering. Here we present the first application of $k$NN summary statistics to observational clustering measurement, using the 1000 richest redMaPPer clusters ($0.1 \leq z \leq 0.3$) from the SDSS DR8 catalog. A clustering signal is defined as a difference in the cumulative distribution functions (CDFs) of $k$NN distances from fixed query points in the observed clusters versus a set of unclustered random points. We find that the $k = 1, 2$-NN CDFs of redMaPPer deviate significantly from the randoms' across scales of 35 to 155 Mpc, which is a robust signature of clustering. In addition to $k$NN, we also measure the two-point correlation function for the same set of redMaPPer clusters versus random points, which shows a noisier and less significant clustering signal within the same radial scales. Quantitatively, the $\chi^2$ distribution for both the $k$NN-CDFs and the two-point correlation function measured on the randoms peak at $\chi^2 \sim 50$ (null hypothesis), whereas the $k$NN-CDFs ($\chi^2 \sim 300$, $p = 1.54 \times 10^{-36}$) pick up a much more significant clustering signal than the two-point function ($\chi^2 \sim 100$, $p = 1.16 \times 10^{-6}$) when measured on redMaPPer. Finally, the measured 3NN and 4NN CDFs deviate from the predicted $k = 3, 4$-NN CDFs assuming an ideal Gaussian field, indicating a non-Gaussian clustering signal for redMaPPer clusters, although its origin might not be cosmological due to observational systematics. Therefore, $k$NN serves as a more sensitive probe of clustering complementary to the two point correlation function, providing a novel approach for constraining cosmology and galaxy–halo connection.

Key words: cosmology: large-scale structure of Universe – galaxies: clusters: general – methods: statistical

1 INTRODUCTION

Galaxy clusters, the largest gravitationally-bound objects in the universe, are also the most strongly spatially clustered set of biased tracers of the cosmic large scale structure (LSS). They have been one of the key testing grounds of the flat-$\Lambda$CDM cosmology model, as well as the nature of dark matter and galaxy formation physics (see Allen et al. 2011; Kravtsov & Borgani 2012 for reviews). Cluster observations in optical (Eisenstein et al. 2011; Rykoff et al. 2016; Abbott et al. 2018), X-ray (Vikhlinin et al. 2009; Mantz et al. 2010, 2016), and millimeter (Bleem et al. 2015; Planck Collaboration et al. 2016b; Hilton et al. 2018) wavelengths, have accumulated a diverse set of cluster properties — e.g., richness, X-ray gas luminosity and temperature, Compton Y in SZ effect — over a wide range of cluster masses. Using the observable-mass scaling relations, one can infer the underlying halo mass function (HMF) which is sensitive to the cosmological model, including the Equation of State (EOS) of dark energy at low redshifts. One successful application of galaxy clusters as a cosmological tool is the constraint on the parameter $\sigma_8(\Omega_m/0.3)^{\alpha}$ via cluster abundances ($\sigma_8$ is the rms linear matter density fluctuation on scales of $8$ Mpc/$h$, and typically $\alpha \sim 0.2 - 0.4$). Notwithstanding, individual cluster masses are hard to measure and this is the main source of systematics that propagates into cosmological parameter estimations using cluster abundances (e.g., weak lensing mass von der Linden et al. 2014; Mantz et al. 2015; Hoekstra et al. 2015). Many recent and ongoing surveys (e.g., SDSS Simet et al. 2017, DES Melchior et al. 2017; McClintock et al. 2019b) have seen improvement in the calibration of these cluster scaling relations, which will become even better with upcoming wide-area surveys (e.g., the Vera Rubin Observatory Ivezić et al. 2019, the Roman Space Telescope Spergel et al. 2015, Euclid Euclid Collaboration et al. 2019, eRosita Predehl et al. 2021).

Apart from cluster abundances, the spatial distribution of clusters as tracers of the most massive dark matter halos also encodes cosmological information. There has been significant effort (Zehavi et al. 2002; Meneux et al. 2008; Gaztañaga...
et al. 2009; Mostek et al. 2013; Tojeiro et al. 2014; Ross et al. 2017) using the two point correlation function and its Fourier transform, the power spectrum, to study the LSS via galaxy clustering. However, as massive \( M_{\text{halo}} \gtrsim 10^{14} M_\odot \) clusters are in the exponential tails of the HMF, their number density is low. The traditional two-point correlation function, despite its many successes as a summary statistics describing galaxy and halo clustering, is dominated by shot noise out to large scales due to the sparsity of these massive galaxy clusters. Furthermore, continued gravitational collapse induces non-Gaussian features in the clustering of these massive objects in the late Universe. A possible way to improve the clustering signal SNR is to gather information of higher-order correlations functions (\( n > 2 \)-point). However, the complexity of the problem increases with \( n \) (\( O(N^n) \), \( N \) being the sample size), making \( n \)-point correlation functions progressively challenging to compute.

An alternative approach is to use a different set of summary statistics which encode information about higher order statistics of the field, while being computationally inexpensive to evaluate on the data. The distribution of \( k \)-nearest neighbor (\( k \text{NN}, \) hereafter) distances have recently been shown (Banerjee & Abel 2021a) to provide an efficient way (\( O(N \log N) \) complexity) of extracting the combined clustering signal of all higher order correlation functions. In the special case of \( k = 1 \), the INN-CDF contains the same information as the Void Probability Function (VPF) introduced in White (1979), which measures the probability that an arbitrarily centered sphere with radius \( R \) will contain no data points.

Given its extra sensitivity to all higher-order correlation functions, while maintaining feasible computational demand, the \( k \text{NN} \) summary statistics provides a much more powerful yet resource-efficient probe of cosmology than the two-point correlation function. On quasi-linear scales, \( k \text{NN} \) distributions of tracers can be modeled using the same set of parameters (Banerjee et al. 2021) that are used to model the two-point correlation function in this regime (e.g. Kokron et al. 2021). Nearest neighbor measurements can also be used to measure higher order cross correlations between two different sets of tracers (Banerjee & Abel 2021b).

In this work, we present the first application of the \( k \text{NN} \) measurements to an observational data set to measure its spatial clustering. We aim to extract clustering information based on 3D distances between a set of query points and the observed data assuming a fixed cosmology. Apart from investigating whether or not a clustering signal can be successfully extracted, we will also compare its sensitivity performance to the two-point correlation function measured on the same data set. Our target data set, the 1000 richest galaxy clusters identified from the SDSS DR8 catalog using the redMaPPer algorithm (Rykoff et al. 2014, 2016), is a volume-limited (\( M_{\text{halo}} \gtrsim 10^{14} M_\odot \)) set of massive clusters which is nearly uniform in comoving number density in the redshift range of \( z \in [0.1, 0.3] \). This sample is almost an order of magnitude higher in mass than previous analyses of galaxy clustering in spectroscopic surveys (e.g., White et al. 2011; Parejko et al. 2013), due to which we expect stronger effects of shot noise, diminishing the clustering signal.

Testing our method on a sparse and potentially shot noise-dominated sample is crucial for verifying whether or not the additional clustering information from \( n > 2 \) point correlation functions in the \( k \text{NN} \) summary statistics actually leads to a more prominent clustering signal compared to the two-point correlation function. Of course, extracting a clustering signal with an assumed cosmology using \( k \text{NN} \) from an observed set of massive clusters is only the first step towards constraining cosmology with this method. Subsequent analysis applying appropriate selection functions, accounting for cluster mass estimate uncertainties, and exploring different constraining methods (emulator versus simulation) in the future are also crucial to go from a clustering signal detection with the \( k \text{NN} \) formalism, as is done here, to a self-contained cosmological constraint analysis using the most massive clusters in the Universe.

The structure of the paper is as follows: Section 2 will introduce the formalism of the \( k \text{NN} \) summary statistics and how clustering is measured with it; Section 3 will briefly overview the redMaPPer cluster sample, the corresponding random sample, and how we set up the random query points for obtaining the \( k \text{NN} \) statistics; Section 4 will present the results of clustering measurements of our selected sample with \( k \text{NN} \) summary statistics compared to two-point correlation functions, quantify and compare their clustering detection significance, and discuss the implications of these results; Section 5 will present the detection of a non-Gaussian signatures in the clustering of the redMaPPer clusters using \( k \text{NN} \) summary statistics; Section 6 will summarize the main conclusions of this study and provide an outlook for future directions of work.

2 METHODOLOGY

We briefly summarize the \( k \text{NN} \) summary statistics formalism introduced in Banerjee & Abel (2021a). The \( k \text{NN} \) formalism describes clustering information in the data (redMaPPer clusters in our case) using their distances from a set of random query points which are distributed over the survey volume. From each query point, one measures the distance to the \( k \)th nearest neighbor data point, \( k = 1, 2, ..., \). Collecting together the distances to the \( k \)th nearest neighbor data point for some specific value of \( k \) from all the random query points, one can construct an empirical cumulative distribution function (CDF) by sorting these distances. This is the summary statistics \( k \text{NN-CDF} \) that we refer to in the following analysis. In contrast, the signal in the two-point correlation function is calculated using distances between pairs of data points.

Mathematically, \( k \text{NN-CDFs} \) are connected to counts-in-cells (CICs) statistics — \( P_{k|V} \), the probability of finding exactly \( k \) data points in a volume \( V \), averaged over all points in the Universe. The generating function of the CICs can be written as (see Appendix A of Banerjee & Abel (2021a) for details):

\[
P(x|V) = \sum_{k=0}^{\infty} P_{k|V} x^k = \exp \left[ \sum_{n=1}^{\infty} \frac{n^N (x-1)^N}{N!} \times F_N \right],
\]

\[
F_N = \int_V \cdots \int_V dr_1 \cdots dr_N \xi^{(N)}(r_1, \ldots, r_N),
\]

where \( x \) is a dummy variable. Given the generating function, it can be seen that \( P(x|V) \) contains all orders of correlation functions \( \xi^{(N)} \) within its functional expansion.

Given the generating function, CIC functions of different
ks can be expressed as:
\[
CIC_k(r) = P_{k|V} = \frac{1}{k!} \left[ \frac{d^k}{dz^k} P(x|V) \right]_{z=0},
\]
where we have assumed that each random cell with volume \( V = 4\pi r^3/3 \) is spherical (we use CIC \( r \) and \( P_{k|V} \) interchangeably in the following). Since the CIC \( r \) is written as a derivative of the generating function in Eq. 1, with respect to the dummy variable \( x \), each CIC \( r \) is sensitive to a different combination of the integrals of all \( N \)-point functions \( f_N \). Historically, the \( k = 0 \) CIC function is known as the ‘Void Probability Function’ defined in White (1979), which describes the probability of finding no points in random cells with volume \( V \).

It is also possible to express the CICs in terms of the cumulative counts:
\[
P_{k|V} = P_{>k-1|V} - P_{>k|V}, \quad \text{for } \forall k \geq 1,
\]
where \( P_{>k|V} \) is the probability of finding more than \( k \) points in the volume \( V \). If the random volume \( V \) is spherical and centered on the query point, the probability \( P_{>k|V} \) is equivalent to the probability of the \( k \)th nearest data point having a distance to the query point smaller than the radius \( r \) of the spherical volume, where \( r = (3V/4\pi)^{1/3} \). Then the probability \( P_{>k|V} \) can be expressed through this equivalence in the CDF value of the \( k \)th nearest neighbor distance at \( r \):
\[
CDF_{(k+1)\text{NN}}(r) = P_{>k|V} = \frac{4}{\pi} r^3 = 1 - \sum_{i=0}^{k} P_{i|V} = \frac{4}{\pi} r^3
\]
where \( CDF_{(k+1)\text{NN}}(r) \) describes the probability of finding the \( k \)th nearest data point within a distance of \( r \) at any given query point. Thus, Equation 3 can be readily converted from volume-based probabilities to distance-based functions:
\[
CIC_k(r) = CDF_{(k-1)\text{NN}}(r) - CDF_{\text{NN}}(r), \quad \text{for } \forall k \geq 1.
\]
This is how \( k \text{NN}-CDFs \) are mathematically connected to CICs. Note that the \( k \text{NN}-CDFs \) are also, therefore, sensitive to different combinations of all the integrated \( N \)-point functions \( f_N \) from Eq. 1.

The distance to the \( k \)th nearest neighbor data point from a set of random query points can be extracted efficiently with the KDTree algorithm. We use the cKDTree function \(^1\) from SciPy in this work. Multiple queries of the same \( k \)th nearest neighbor distances of different \( k \)s fast (time complexity \( \mathcal{O}(N \log N) \)) once the 3D spatial tree is constructed for a specific set of data points. We will describe in more detail in Section 3.3 how we set up random query points for the SDSS redMaPPer cluster catalog.

3 THE SDSS-DR8 REDMAPPER CLUSTER CATALOG

3.1 Observed clusters

As mentioned in the previous sections, our main goal is to test the methodology of \( k \text{NN} \) summary statistics on a set of sparse but clustered data points. Large galaxy cluster catalogs have been established observationally in recent years thanks to concerted efforts in X-ray (e.g., Chandra Vikhlinin et al. 2009), SZ (Sunyaev-Zel’dovich effect e.g., Planck Planck Collaboration et al. 2014b, 2016b, SPT Bleem et al. 2015, and ACT Hilton et al. 2021), and optical (e.g., SDSS Eisenstein et al. 2011) surveys. Out of these existing data sets, we choose the redMaPPer cluster catalog (Rykoff et al. 2014, 2016) identified from the SDSS DR8 catalogs. The advantage of redMaPPer clusters for our purpose is two-fold: i) the redMaPPer cluster sample is almost complete and uniform in comoving number density for the massive clusters we are interested in at low redshift \( (z \lesssim 0.3) \), ii) there already exists a large mock catalog of random points processed through the redMaPPer pipeline that accounts for all selection functions inherent to the catalog. Therefore, the well-studied redMaPPer cluster catalog serves as a neat starting point for our analysis. We use the latest version (v6.3) of the publicly available redMaPPer data products for the following analysis \(^2\).

The redMaPPer (Rykoff et al. 2014; Rozo & Rykoff 2014; Rozo et al. 2015a,b) cluster finding algorithm determines cluster membership of galaxies projected around galaxy clusters in a probabilistic manner. The model infers the photometric redshift (\( z_\lambda \)), cluster richness (\( \lambda \)), and cluster center location jointly, which is achieved by training on a set of red-sequence galaxies with spectroscopic redshift selected from the SDSS (Aihara et al. 2011; Eisenstein et al. 2011) and GAMA (Driver et al. 2011) surveys. Hence, each cluster is assigned a probability density distribution of their photometric redshift as predicted by the model. Since the photometric redshift errors \((|\Delta z/(1 + z)| \approx 0.006, \text{ Rykoff et al.})\)
are small for the redMaPPer clusters, we choose their most-likely photometric redshift as the fiducial redshift for calculating distances. We convert cluster redshifts into comoving distances assuming a flat-ΛCDM cosmological model with Planck 2015 parameters (Planck Collaboration et al. 2016a): $h = 0.6774$, $\Omega_m = 0.3089$, $\Omega_{\Lambda} = 0.6911$. We have verified that the 1NN and 2NN CDFs measured in the following vary by $\lesssim 3\%$ across the chosen radial ranges (see Section 3.3) if we sample the posterior of the inferred photometric redshift for each cluster. Therefore, our following analysis will not be significantly affected by the photometric redshift uncertainties.

Following previous work (e.g., Rykoff et al. 2016; To et al. 2020), we first select clusters within the redshift range of $0.1 \leq z \leq 0.3$. The upper redshift limit corresponds to where the galaxy luminosity threshold adopted by redMaPPer goes below the SDSS survey depth. The lower redshift limit is mainly set by the volume limit for massive clusters in the very local universe. Following the SDSS survey footprint, the redMaPPer clusters that fall within $0.1 \leq z \leq 0.3$ make up a sample not only uniform in RA/DEC but also in comoving number density (see Fig. 18 in Rykoff et al. 2014). Out of the 6647 clusters in this redshift range, we pick out the 1000 richest clusters as our "data" sample in our following analysis. This sample size choice is motivated by our expectation that $kNN$ summary statistics can potentially detect a clustering signal on this sparse dataset where the two-point correlation function would be dominated by shot noise. This selected sample has a minimum richness of $\lambda_{\text{min}} \approx 44$ which is $\geq 98\%$ complete in $0.1 \leq z \leq 0.3$ (see Fig. 22 in Rykoff et al. 2014). Moreover, this selected cluster sample is also massive, with a halo mass threshold of a few times $10^{14} M_{\odot}$ (see recent constraints e.g., Simet et al. 2017; Murata et al. 2018; $\lambda > 20$ corresponds to $M \geq 10^{14} M_{\odot}$ for $z < 0.35$, Rykoff et al. 2014). The average comoving number density of our sample is $5.26 \times 10^{-7} \text{Mpc}^{-3}$. The redshift and richness distribution of our final 1000-cluster sample is shown in Fig. 1. The RA and DEC of the 1000 selected clusters are shown in the upper panel of Fig. 2.

### 3.2 The random mock cluster sample

In order to quantify the significance of the clustering signal for our selected sample using either $kNN$ summary statistics or two-point correlation functions, we need to compare the measurements to a sample of random points with no underlying clustering processed through the observation pipeline of the redMaPPer clusters. We use the mock observation of random points from publicly available ‘random’ catalog in redMaPPer data products. This will largely remove systematic biases in the selection functions of the "data" sample, and any remaining differences in the clustering measurements between the data and random samples will be due to clustering.

As described in Rykoff et al. (2014), random RA and DEC of the mock clusters are first generated that samples the survey mask. Then, their mock richness and photometric redshifts are drawn from the corresponding distributions of the observed redMaPPer clusters. Given these quantities, 5000 mock galaxy members in each cluster are generated using Monte Carlo sampling of a projected NFW density profile (with appropriate smoothing applied), which is then sub-sampled to match the mock richness generated in the previous step. In the end, every mock cluster is placed at 100 random directions on the sky, virtually "observed" with the redMaPPer cluster finder, and processed with the corresponding bright stars and survey edges masks. The resulting random catalog is ideal for large scale structure studies like this paper as it characterizes the "noise level" of the data set from which we are trying to measure clustering. Any deviation beyond this "noise level" would indicate the presence of clustering in the data set.

Since the random catalog is densely sampled, we end up with 124672 mock clusters after applying the redshift ($0.1 \leq z \leq 0.3$) and richness ($\lambda_{\text{min}} \geq 44$) cuts to match our 1000 selected redMaPPer "data" sample. From these 124672 mock clusters, we randomly sub-sample 1000 of them to create a "random" sample to which we compare the clusters (data sample). One realization of a "random" sample is shown in the bottom panel in Fig. 2. As expected, the random sample are visually less clustered (more diffuse clumps, smaller voids) compared to the redMaPPer clusters in the top panel.

### 3.3 Setting up the query points

In order to obtain the $kNN$ summary statistics for our samples, we also have to create an appropriate set of query points from which we measure the $kNN$ distances to the data/random clusters. In the following, we refer to comoving distances if not specified otherwise. We start by placing an equi-distant 3D grid of $100 \times 100 \times 100 = 10^6$ query points

---

**Figure 2.** *Upper panel:* The RA/DEC distribution of the 1000 selected redMaPPer clusters. The redshift range is $0.1 \leq z \leq 0.3$ while the minimum richness is $\lambda_{\text{min}} \approx 44$. *Lower panel:* The RA/DEC distribution of a sample drawn from the random catalog processed through the redMaPPer selection pipeline. The random sample shares the same sample size, redshift range, and minimum richness cut as the data in the upper panel.
Figure 3. Top left: The peaked-CDFs (CDF(r)) on scales where CDF(r) ≤ 0.5, and 1 – CDF(r) on scales where CDF(r) > 0.5) for the k = 1, 2 nearest neighbor distributions. Top right: The P_k|V (CICs) as a function of radial scale (volume) for the k = 1, 2 nearest neighbor distributions. In the top panels, the solid curves represent the 1000 richest redMaPPer clusters (data), while the shaded regions with the same color as the solid curves represent the 1,2,3 σ scatters for 2000 realizations of random samples, each having 1000 mock clusters.

(the choice is empirical as long as it is dense enough to create a smooth CDF) in a cubic volume centered at z = 0 with side length 2 × 1231.86 Mpc, where 1231.86 Mpc corresponds to the distance at the maximum cluster redshift z = 0.3. Next we select the query points that lie within the spherical shell containing the sample's redshift range 0.1 ≤ z ≤ 0.3 (distance ∈ [432.13, 1231.86] Mpc). Out of the query points in this shell, we select those having RA and DEC that follow the SDSS Survey mask (≈ 10^4 deg^2), i.e. query points projected into Nside = 2048 healpy^3 pixels with at least half of each heal-pixel's sky area being in the survey footprint (f_good > 0.5). We also require the query points to lie in a heal-pixel where the survey depth is complete to z = 0.3 (z_max ≈ 0.3). We end up with a total of 123386 query points that densely sample the survey volume occupied by the data/random clusters, from which we measure both the k = 1 and 2-NN distances. In Appendix D, we show a visualization (Fig. D1) of a set of query points at a fixed redshift (not the same as the 3D query points introduced above which we use for our following analysis) that follow the SDSS survey mask colored by their kNN distances to the 1000 richest redMaPPer clusters for a more intuitive illustration of the distance query procedure.

The observed/mock clusters are assumed to be point sources located at the cluster centers. We construct KDTrees using the locations of the observed/mock clusters and conduct distance query from the query points as mentioned above. We store the distances from the query points to their kth-nearest data (observed) or random (mock) point in an array. Once we line up all these kNN distances in ascending order, we can construct an empirical CDF with a resolution of ≈ 0.001 (we linearly interpolate the CDFs in between each measured kNN distance). To ensure that the CDF functions are not over-extrapolated into the tails where there might be very few data points, we limit our CDF measurements for each k to the scales where that specific CDF has values in the range of ≈ [0.05, 0.95]. This results in a set of adopted radial ranges for different ks:

- k = 1: r ∈ [35/Mpc, 135/Mpc], 25 logarithmic bins.
- k = 2: r ∈ [55/Mpc, 155/Mpc], 25 logarithmic bins.

Using Equation 5, we can then convert the measured kNN CDFs to the corresponding kth order CICs for the data/random sample (note that we measure the k = 3 CDF in addition to k = 1, 2 for calculating the CIC functions of order k = 2 in the radial range from 55 to 155 Mpc).

4 RESULTS

4.1 CDF and P_k|V measured with kNN distances

In this section, we present the CDF and P_k|V measured for k = 1, 2 nearest neighbor distributions of the 1000 richest redMaPPer clusters compared to the mock random samples. We choose to show the peaked CDF (pCDF(r)) instead of the conventional CDF(r) to better elucidate the features in both tails of the CDF. The pCDF is defined as CDF(r) on scales where CDF(r) ≤ 0.5, and 1 – CDF(r) on scales where CDF(r) > 0.5, which .

The results for the peaked CDFs (left panel) and P_k|V (CICs, right panel) functions of the 1000 richest redMaPPer clusters along with that of the 2000 random samples are shown in Fig. 3. In both panels, the solid curve stands for redMaPPer while the colored regions indicate the 1,2,3 σ regions for the distribution of the 2000 random realizations. It can be seen that both the CDFs and CICs of the redMaPPer clusters deviate significantly (≥ 2σ) from the 2000 randoms on almost all scales probed by the k = 1, 2 nearest-neighbor distances. For each individual k, the deviations are scale dependent, with crossings between the redMaPPer CDFs and the randoms’ CDFs for k = 2. The clear and statistically significant difference in the kNN-CDFs (and P_k|V) of the data and the random samples indicate a robust detection of the underlying clustering of the data, while accounting for the

3. https://healpy.readthedocs.io/en/latest/
inherent Poisson noise of a non-clustered sample. We explore the significance of the detection further in next few sections, but it is worth reiterating here that this is the first successful application of kNN summary statistics on an observational sample.

To understand the different clustering information contained in the 1NN and 2NN CDFs, we show the ratio of the redMaPPer clusters’ CDFs to the 2000 random samples’ CDFs in Fig. 4. In each panel, the mean of the 2000 randoms’ CDF is shown as a constant horizontal line (y = 1). Around that line, a gray shaded region denotes the 1σ (68% scatter) distribution of the individual random CDF relative to their mean. The colored solid curve in each panel is the redMaPPer clusters’ CDF relative to the mean of the 2000 randoms’ CDF, while the shaded region around the solid curves show the 1σ Jackknife errors. We obtain the Jackknife errors using the leave-one-out method. Particularly, for each Jackknife sample, we leave out the query points having k = 1 or 2 NN distances that are associated with one specific cluster, forming one thousand 999-cluster samples. In each Jackknife sample, this procedure effectively removes a part of the survey volume that contains the left-out cluster and query points in its vicinity. This ensures that the averaged kNN CDFs of all Jackknife samples remain consistent with the measurements on the 1000 richest redMaPPer clusters. As we show in Appendix A, the leave-one-out Jackknife error is overall a good estimate of the realization-to-realization (intrinsic) kNN-CDFs’ scatter when tested on cosmological N-body simulations, although on large scales it can slightly underestimate the intrinsic scatter. We leave a detailed comparison of different error estimation methods on the measurement errors of kNN CDFs to future work.

From Fig. 4, it is evident that the k = 1 relative CDF differs from the randoms in a different manner compared to the k = 2 CDF. For k = 1, the redMaPPer CDF is significantly suppressed at all scales compared to the randoms’ CDF, with the largest suppression occurring at r ~ 70 Mpc. Compared to the randoms, the redMaPPer clusters will clump in higher density regions and create larger cosmic voids, leading to higher probabilities for query points to be located in these voids. When the majority of query points live in these underdense regions, their closest cluster neighbor, i.e. the 1NN, will be farther away to the query points on average as compared to the case where the nearest-neighbors are unclustered and randomly-distributed. The net effect is a significant shift of the 1NN-CDF for redMaPPer clusters to larger scales with respect to the randoms’, resulting in a smaller value for the redMaPPer 1NN-CDF when evaluated at the same radial scale as the randoms. There will be a small fraction of query points living in overdense cluster regions contributing to the smallest scales in the CDF, but that will be submerged by Poisson noise due to small number statistics, as random points can also form small patches of overdense regions randomly.

For k = 2, the differences of the redMaPPer CDF relative to the randoms’ CDFs are mainly expressed as an enhancement at small scales and suppression at large scales. In the case of 1NN which traces query-data point separation, strong clustering leads to an increase in the distance from a query point to its nearest cluster neighbor, which shifts the 1NN-CDF to large scales. The 2NN distance adds another layer of data-data separation on top of the query-data separation of the 1NN distance. At small scales, once the 1NN cluster of a query point is found, the 2NN cluster will tend to be closer to the 1NN due to clustering, leading to smaller 2NN distances from the clusters to the query point, tracing out overdense regions of the clusters. This scenario applies to query points living in both under and overdense regions who have their
1NN and 2NN clusters in the same direction, both contributing to the small scales of the 2NN CDF. This then shifts the 2NN CDF to smaller scales compared to the randoms, and create an enhancement of the relative 2NN CDF at scales of \( r \lesssim 80 \) Mpc. However, we also observe a suppression of the redMaPPer 2NN CDF with respect to the randoms’ at larger scales. This is mainly contributed by query points living in voids having their 1NN and 2NN clusters on opposite sides of the void. The 2NN distances will increase for these query points under stronger clustering which creates larger voids, eventually driving the 2NN CDF to larger scales compared to the randoms. The cross over point between the CDFs of the clusters and randoms mark the average intra-cluster distance, transitioning from overdense to underdense regions. This feature is also shown in the left panel of Fig. 3 where the redMaPPer 2NN CDF cross over point coincides with where the redMaPPer 1NN CDF reaches 0.5.

### 4.2 Two-point correlation functions

In this section, we present the clustering signal measured from the two-point correlation function on the same data set in order to contrast the clustering significance detected by the \( kNN \) summary statistics shown in the previous section.

For the two-point correlation function, we choose the commonly used Landy & Szalay (1993) estimator:

\[
\xi(r) = \frac{DD(r) - 2DR(r) + RR(r)}{RR(r)},
\]

where \( DD \) stands for the data cross data, \( DR \) stands for data cross random, and \( RR \) for random cross random terms, respectively. The \( D \) term in our measurements is either the 1000 richest redMaPPer clusters or one of the mock random cluster samples. We fix the \( R \) terms to be measured from a set of \( 1 \times 10^5 \) random points (100 times in number density) drawn from the parent sample of all 124672 mock clusters processed through the redMaPPer selection functions. The radial range over which we measure \( \xi(r) \) is set to \( r \in [35 \text{ Mpc}, 155 \text{ Mpc}] \) with 50 logarithmic bins, matching the radial range and number of bins covered by the \( k = 1, 2 \)-NN CDFs. Finally, the two point correlation function calculations were done using the public PYTHON package corrfunc \(^4\).

In Fig. 5, we show the two-point correlation functions \( \xi(r) \) for the 1000 richest redMaPPer clusters (red) compared to the 2000 random samples (blue). The top panel shows the \( \xi(r) \) measurements, while the bottom panel shows the relative \( \xi(r) \) of the redMaPPer clusters scaled to the mean and 68% scatter of the 2000 randoms. The clustering signal picked up by the two-point correlation function is very noisy compared to the \( kNN \) CDFs. Comparing the redMaPPer clusters to the randoms, the clustering signal for redMaPPer is consistent with the randoms within \( 1\sigma \) confidence interval, considering the intrinsic scatter of the randoms and the Jackknife errors (Norberg et al. 2009, also see Section A for more details) of the observed clusters. The redMaPPer \( \xi(r) \) do seem to demonstrate a more significant clustering signal towards small scales \( (r \lesssim 50 \text{ Mpc}) \), although the Jackknife errors are also quite large. Referring back to Fig. 4, the \( k = 1, 2 \) NN CDFs on the other hand exhibit robust clustering signals for clustering at all scales within the same radial range compared to the randoms, especially at the large scales \( (r \gtrsim 100 \text{ Mpc}) \). This comparison highlights the advantages of \( kNN \) CDFs at detecting clustering signal over the two point correlation function, and we will provide a more quantitative comparison of the clustering signal detection via the difference in \( \chi^2 \) of these two methods in Section 4.4.

---

\(^4\) https://corrfunc.readthedocs.io/en/master/

---

Figure 5. **Upper panel:** The two point correlation function \( \xi(r) \) for the 1000 richest redMaPPer clusters (red) and 2000 realizations from the random catalog (blue). The radial range and number of radial bins are the same as the CDF measurements in Figs. 3 and 4. The red diamond markers denote the \( \xi(r) \) measurement for the redMaPPer clusters, the red shaded region denotes the Jackknife errors (1\( \sigma \)) derived from 200 Jackknife sub-samples of the 1000 redMaPPer clusters (Appendix A). The dashed blue curve shows the mean of the 2000 random samples’ mean \( \xi(r) \), while the shaded blue shaded region denotes the 68% scatter of the random samples. **Lower panel:** The deviation of the redMaPPer clusters’ \( \xi(r) \) from the mean of the 2000 random samples’ \( \xi(r) \) in units of the \( \xi(r) \) distribution standard deviation of the 2000 random samples. Notice that \( \xi(r) \) for redMaPPer and the randoms become undifferentiated (within 1\( \sigma \)) at scales of \( r \gtrsim 100 \text{ Mpc} \), while \( kNN \) CDFs still detect robust clustering signals at these large scales (Fig. 4).
Note that the outer product implies that estimate the covariance matrix, we apply the Hartlap correction factor (Hartlap et al. 2007) to get an unbiased estimate of the inverse covariance matrix:

\[ C^{-1} = \frac{n - N - 2}{n - 1} (c')^{-1}. \] (9)

In Fig. 6, we show the correlation matrix of the joint \( k = 1, 2 \) NN CDFs in the left panel and of the two point correlation function \( \xi(r) \) in the right panel. The correlation matrix \( \mathcal{R} \) is defined as:

\[ \mathcal{R}_{ij} = \frac{C'_{ij}}{\sqrt{C'_{ii}C'_{jj}}}, \] (10)

which is a normalized version of the covariance matrix and has its elements being the auto and cross correlation coefficients (we use covariance and correlation matrices interchangeably in the following). The radial ranges for the \( x \) and \( y \) axes in the two plots were defined in Section 3.3 and 4.2 (also see Figs. 3 and 5).

For the randoms’ joint \( k \) NN CDFs, the covariance matrix exhibits pronounced auto and cross-correlation behaviors across different \( k \)s. Each pair of auto or cross correlations also vary with the radial scale distinctively as being probed by the different CDFs. The distinct features of the auto and cross-correlations between different \( k \)s and radial ranges indicate how the different measurements of the summary statistics respond to the underlying change in local clustering strength caused by Poisson fluctuations in the randoms. The key to understanding these variations in the randoms’ \( k \) NN CDFs is that the overall number density of the data points (here data point refers to the location of a mock cluster) is kept constant across different realizations of the random clusters.

For the \( k = 1 \) auto-correlation block, if random fluctuations lead to slightly stronger clustering, more query points will live in underdense voids and shift the 1NN CDF to larger scales. This will result in a decrease of the 1NN CDF at fixed radial scales when there is more clustering, leading to the apparent positive correlation within the block. The small scales of the 1NN CDF is probed by a rare portion of query points living in overdense regions that are very close to the mock random

Figure 6. Left panel: The correlation matrix (defined in Eq. 10) of the combined \( k = 1, 2 \) NN CDFs for the 2000 random samples (25 radial bins for each \( k \), 50 bins combined). The radial range for the 1NN and 2NN CDFs are \( r \in [35 \, \text{Mpc}, 135 \, \text{Mpc}] \) and \( r \in [55 \, \text{Mpc}, 155 \, \text{Mpc}] \), respectively. Right panel: The correlation matrix of the two point correlation function \( \xi(r) \) for the 2000 random (50 radial bins). The radial range for \( \xi(r) \) is \( r \in [35 \, \text{Mpc}, 155 \, \text{Mpc}] \). The correlation matrix of the \( k \) NN CDFs show significant (anti-)correlation between different scales and different \( k \)s, whereas \( \xi(r) \) shows almost no correlation between different radial scales.
clusters. Shuffling the location of the mock clusters randomly in a slightly overdense region will have a minimal effect on the distances between these points and the majority of query points living in voids, hence the correlation of the 1NN CDF at small and large scales is weak.

For the $k = 2$ auto-correlation block, the small and large scales are anti-correlated due to different responses of the CDF at these scales. If the clustering strength slightly increases due to random fluctuations, the $k = 2$ NN distances from a query point to its mock cluster neighbor could have two different scenarios. The small scale tail of the 2NN CDF mainly probes mock clusters in overdense regions, which shifts the 2NN CDF to smaller scales with increased clustering. On the other hand, the number of massive clusters living in a single overdensity is always limited, and the 2NN for the query point might be across a void (an underdense region). Hence, the large scale tail of the 2NN CDF is dominated by these cases and will shift to larger scales in the event of increased clustering due to random fluctuations. Hence, for the 2NN auto-correlations, the small and large scales of the 2NN CDFs will respond oppositely in the event of increased clustering due to random fluctuations, leading to the apparent anti-correlation seen in Fig. 6.

Understanding the auto-correlations of the $k = 1, 2$-NN CDFs also exemplifies their cross-correlation behavior. For the cross-correlation between the 1NN and 2NN CDFs, the shift of the 1NN CDF is almost monotonic towards larger scales if clustering strength increases, while the 2NN CDF respond oppositely at small and large scales. These auto and cross correlation features manifested by the joint $k$NN CDFs’ covariance matrix are also consistent with the comparison between the redMaPPer and random clusters’ CDFs shown in Fig. 3, although the clustering strength difference there is much larger than the internal fluctuations among the randoms.

Last but not least, the covariance matrix of the $\xi(r)$ measurements for the 2000 random samples is shown in the right panel of Fig. 6. The absolute value of all off-diagonal terms are tiny compared to the diagonal terms. This indicates that the two-point correlation function measurements do not show any coordinated scale-dependent response to random fluctuations in the clustering strength of the mock data points. This is in stark contrast with the $k$NN CDFs as the latter is intrinsically correlated on different scales even in the absence of clustering.

4.4 Distribution of $\chi^2$ values

Having understood the intrinsic properties of the two summary statistics when applied to the randoms in the previous section, we will present in this section the distribution of $\chi^2$ values for the 1000 richest redMaPPer clusters with respect to the 2000 random samples measured by the joint $k$NN CDFs and the two-point correlation function, respectively. This will further quantify the difference in the statistical power of these two summary statistics at extracting spatial clustering signals from the redMaPPer clusters.

Given the covariance matrix and its (unbiased) inverse for the randoms, the deviation of any individual data vector (i.e. joint $k$NN CDFs or $\xi(r)$ of the redMaPPer clusters) from the distribution of the randoms can be expressed in terms of:

$$\chi^2 = (\mathbf{d} - \bar{\mathbf{r}})^T \mathbf{C}^{-1} (\mathbf{d} - \bar{\mathbf{r}}) = \sum_{ij} C_{ij}^{-1} (d_i - \bar{r}_i)(d_j - \bar{r}_j),$$

where $\mathbf{d}$ is the corresponding data vector ($k$NN CDFs or $\xi(r)$). Alternatively, we could replace $\mathbf{d}$ with the $i$-th random vector to obtain the $\chi^2$ for the $i$-th random sample with respect to its own distribution. This quantity reflects the combined statistical deviation of $\mathbf{d}$ from $\bar{\mathbf{r}}$ accounting for the intrinsic scatter of the randoms at all relevant radial scales. A larger $\chi^2$ would indicate a more significant deviation of the data vector from the randoms, indicating a more significant clustering signal detection.

The results of the $\chi^2$ values for the joint $k = 1, 2$-NN CDFs (top panel) and the two point correlation function (bottom panel) are shown in the Fig. 7. The $\chi^2$ distribution of the randoms are almost identical in both panels as a result of the aligned radial bins applied to the $k$NN CDFs $\xi(r)$. In both panels, $\chi^2$ for the 1000 richest redMaPPer clusters do not overlap with the $\chi^2$ of the randoms (peaking at $\chi^2 \sim 50$), indicating that both summary statistics can detect a clustering signal for the redMaPPer clusters that is not overwhelmed by shot noise. However, the $\chi^2$ of the joint $k = 1, 2$-NN CDFs ($\chi^2 = 293.33$) is much larger than that of $\xi(r)$ ($\chi^2 = 110.63$), which translates to $p$-values of $1.54 \times 10^{-36}$ ($k$NN) and $1.16 \times 10^{-16}$ ($\xi$) assuming that the null hypothesis follows a $\chi^2$ distribution with 49 Degrees of Freedom (DoF, number of radial bins – 1). Since the different radial scales are correlated for the $k = 1, 2$-NN CDFs (Fig. 6 left panel), the effective DoF will be less than 49 and the above $\chi^2$ estimate serves as a lower bound for the $k$NN CDF clustering detection. These findings quantitatively justify that the $k$NN summary statistics is a more sensitive probe of clustering for the 1000 richest redMaPPer clusters compared to the two-point correlation function. As we will discuss in the following, the two point correlation function is more sensitive to close pairs at small scales (see Section 4.5 and Appendix C), while the sensitivity gain of the $k$NN CDFs arise from the ability to probe non-Gaussian clustering information that is not captured by the two point correlation function (see Section 5).

4.5 Discussion

As shown above, we have compared the $k$NN CDFs with the two-point correlation function measurements on the 1000 richest redMaPPer clusters, with the $k$NN CDFs detecting a more robust clustering signal. To verify that the CDFs measured for the redMaPPer clusters are reasonable, we show the theoretical predictions of the $k = 1, 2$ CDFs for the 526 most massive clusters in a $1(Gpc/h)^3$-box dark-matter-only cosmological simulation (the QUIJOTE suite, Villaescusa-Navarro et al. 2020) in Appendix B. The choice of the sample size aims to match the number density of the 1000 richest redMaPPer clusters. Comparing Fig. B1 with Fig. 4, the shape of the relative CDFs as well as the Jackknife errors for the 1000 richest redMaPPer clusters are in good agreement with the 526 most massive clusters in the simulation. The availability of multiple realizations at a fixed cosmology in the QUIJOTE simulations also allow us to investigate the intrinsic scatter in the clustering expected from a sample with the number density of the redMaPPer clusters in the data.
Figure 7. Upper panel: $\chi^2$ distribution for the joint $k = 1, 2$ nearest neighbor CDFs, with the blue histograms showing the 2000 random samples and the dashed vertical red line showing the $\chi^2$ of the 1000 richest redMaPPer clusters. Lower panel: $\chi^2$ distribution for the two point correlation function $\xi(r)$, with the blue histograms showing the 2000 random samples and the dashed vertical red line denoting the $\chi^2$ of the 1000 richest redMaPPer clusters. The $\chi^2$ distribution for the randoms in the top and bottom panels are almost identical, serving as the null hypotheses for the clustering detection of the two summary statistics.

To this end, we use 1000 different realizations at the fiducial cosmology of the QUIJOTE suite. Denoted by the error bars in Fig. B1, the intrinsic scatter due to random draws are smaller than the Jackknife errors for both 1NN and 2NN CDFs across the entire radial range, suggesting that the Jackknife errors are more conservative and do not underestimate the expected errors in the clustering signal detection.

Another note to highlight is on the scale-dependent sensitivity of the two-point correlation function. The $\xi(r)$ measurement shown in Fig. 5 is noisy but still demonstrates an increasingly significant signal towards smaller radial scales. This indicates that $\xi(r)$ is more sensitive to close pairs of clusters in overdense regions with small separations. We show in Appendix C that the clustering signal detection from $\xi(r)$ becomes more significant if we extend the radial range lower limit down to 20 Mpc (Fig. C1). Although that signal of $\chi^2 \sim 250$ is not as significant as the joint $k$NN CDFs, it shows where the sensitivity of the two-point correlation function originates from. As the $k = 1, 2$-NN CDFs have sensitive clustering detection on scales $\sim 100$ Mpc, we suggest that the $k$NN formalism is a more sensitive and scale-complementary probe of large scale structure clustering as compared to the two-point correlation function.

5 PROBING NON-GAUSSIANITY

As shown in Section 4.4, the $k$NN CDFs have a much higher sensitivity of clustering compared to the two-point correlation function, which could be due to the different scopes of information probed by these summary statistics. Since the two-point correlation function only fully describes a Gaussian density field, any non-Gaussian feature of the density field would not be captured by $\xi(r)$. However, the $k$NN CDFs effectively measure the combined clustering signal from all orders of correlation functions and has the capability of detecting beyond-Gaussian clustering signals, which could improve its sensitivity over $\xi(r)$. Therefore in this section, we evaluate the significance of the non-Gaussian clustering signal measured by the $k$NN CDFs on the 1000 richest redMaPPer clusters relative to the predicted CDFs assuming an ideal Gaussian field.
To quantify the significance of non-Gaussian clustering signals picked up by the $k = 1, 2$-NN CDFs, we calculate the 3NN and 4NN CDFs predicted from the measured 1NN and 2NN CDFs assuming a Gaussian random field, and compare them to the measured 3NN and 4NN CDFs of the 1000 richest redMaPPer clusters. Using Equations 17 to 21 in Banerjee & Abel (2021a), the predicted 3NN and 4NN CDFs using the 1NN and 2NN CDFs in an ideal Gaussian field can be expressed as (we drop the variable $r$ for brevity):

$$\text{CDF}_{4NN} = \text{CDF}_{2NN} + \left[ (1 - \text{CDF}_{1NN}) \log (1 - \text{CDF}_{1NN}) \right. $$

$$+ \left( \text{CDF}_{1NN} - \text{CDF}_{2NN} \right) $$

$$- \frac{1}{2} \left( \text{CDF}_{1NN} - \text{CDF}_{2NN} \right)^2 $$

$$
\text{CDF}_{3NN} = \text{CDF}_{2NN} + \left[ \left( \text{CDF}_{1NN} - \text{CDF}_{2NN} \right) \log (1 - \text{CDF}_{1NN}) \right. $$

$$+ \left( \text{CDF}_{1NN} - \text{CDF}_{2NN} \right)^2 $$

$$ - \frac{1}{6} \left( \text{CDF}_{1NN} - \text{CDF}_{2NN} \right)^3 \right]. \tag{12}$$

In the top panel of Fig. 8, we show the measured $k = 1, 2, 3, 4$-NN CDFs for the redMaPPer clusters (solid curves), along with the 3NN(1,2) and 4NN(1,2) CDFs predicted using the measured 1NN and 2NN CDFs assuming a Gaussian field (dotted curves). The shaded regions around the solid curves denote the 1σ Jackknife errors estimated from 1000 leave-one-out Jackknife samples of the $k = 1, 2, 3, 4$-NN CDF measurements. The predicted 3NN and 4NN CDFs deviate significantly from the measured 3NN and 4NN CDFs, and the Jackknife errors cannot fully account for the apparent offsets. This suggests that the clustering of the data as measured by the nearest neighbor distributions are not statistically consistent with that of tracers of a perfectly Gaussian random field.

To further quantify this deviation, we compare the $\chi^2$ distribution of the joint $k = 3, 4$-NN CDFs measurements to the predicted joint $k = 3, 4$-NN CDFs assuming that the clusters trace an underlying Gaussian density field. The latter is obtained by propagating the 1NN and 2NN CDF measurements through Equations 12 and 13.

5 Since these two equations are non-linear with the $k = 1, 2$-NN CDFs measurements, Jackknife errors on the 1NN and 2NN CDFs do not translate into the Jackknife errors of the predicted 3NN and 4NN CDFs. This is why we could not quote error bars on the predicted 3NN and 4NN CDFs, and instead calculate the covariance matrix on the 1000 measured Jackknife samples of the $k = 3, 4$-NN CDFs.
the prediction to avoid division by 0 in the \((1 - \text{CDF}_{\text{NN}})\) denominators in Equations 12 and 13 where the 1NN CDF approaches 1 at \(r \gtrsim 200\) Mpc.

The results of \(\chi^2\) are shown in the bottom panel of Fig. 8. Compared to the measurements of the \(k = 3, 4\)-NN CDFs \((\chi^2\text{ peaked at } \sim 40)\), the joint \(k = 3, 4\)-NN CDFs’ prediction have a very significant \((\chi^2 \sim 520)\) deviation. This is a robust detection of non-Gaussian clustering in the 1000 richest redMaPPer clusters, which is sensitively and efficiently probed by the \(k\)-NN CDFs. This also confirms that the extra sensitivity of clustering in the \(k\)-NN-CDFs arise from the unique capability of probing non-Gaussian clustering signal compared to the two-point correlation function. However, we caution that this non-Gaussian signal does not necessarily indicate a robust cosmological origin, since observational systematics such as large scale brightness fluctuations and survey mask features (volume geometry and holes of masked out spaxels) in SDSS might fold into the detected non-Gaussian signal. We perform the same analysis on simulated clusters from the \textsc{Quijote} simulations in Appendix B, and indeed the non-Gaussian signal’s \(\chi^2\) reduces significantly on this mock sample free from observational systematics. Therefore, further investigation in various observational systematics is needed to better understand the cosmological origin of the non-Gaussian clustering probed by the \(k\)-NN CDFs.

6 CONCLUSIONS AND OUTLOOK

In this paper, we have successfully applied the \(k\)-NN summary statistics to the 1000 richest redMaPPer clusters and measured a robust clustering signal. Specifically, we measured the \(k = 1, 2\) nearest-neighbor distances from \(\sim 10^5\) volume-filling query points that densely trace the redMaPPer clusters in the SDSS survey volume at redshifts of \(0.1 \leq z \leq 0.3\). This is the first successful application of the \(k\)-NN formalism on a large observational data set, and paves a broad avenue for future large scale structure studies using this method. Our main findings can be summarized as follows:

- The CDF/CIC functions of the observed clusters show significant deviation from that of 2000 random samples of mock clusters over the radial scales of \(r \in [35\text{ Mpc}, 155\text{ Mpc}]\) (Fig. 3).

- In the presence of clustering, the 1NN CDF of the redMaPPer clusters is suppressed in our radial scale range compared to the randoms due to its sensitivity to cosmic voids. The 2NN CDF show enhancement at small scales by tracing overdense regions, and suppression at large scales by tracing underdense regions (Fig. 4).

- The two-point correlation function \(\xi(r)\) also measures a more noisy and less significant clustering signal compared to the \(k\)-NN CDFs. Nonetheless, there is an increasing trend of the clustering signal towards small scales at \(r \lesssim 50\) Mpc. (Fig. 5).

- We showed the correlation (covariance) matrices of the joint \(k\)-NN CDFs and \(\xi(r)\) on the 2000 random samples in Fig. 6. The \(k\)-NN CDF formalism show strong scale-dependent (anti-)correlations while the two-point correlation function is almost decorrelated at different scales. This is a distinctive difference between the two summary statistics.

- The distribution of \(\chi^2\) of the joint \(k = 1, 2\)-NN CDFs \((\chi^2 \sim 300)\) for the redMaPPer clusters is significantly larger than that of the two-point correlation function \((\chi^2 \sim 100)\), indicating a more robust detection of clustering by the \(k\)-NN CDFs Fig. 7.

- A non-Gaussian clustering signal is detected for the redMaPPer clusters, with a deviation of the predicted 3NN and 4NN CDFs assuming an ideal Gaussian field from the measured ones (Fig. 8). This suggests that the extra clustering sensitivity over the two-point correlation function arises from \(k\)-NN’s ability to capture non-Gaussian clustering information efficiently. However, the origin of this signal might not be fully cosmological and further studies are required to better understand the effects of observational systematics.

The success of extracting a clustering signal using the \(k\)-NN summary statistics on the 1000 richest redMaPPer clusters highlight its potential as a complementary probe of clustering to the traditional two-point correlation function, which is especially sensitive to clustering at large scales and non-Gaussian information. However, it is only a first step towards making quantitative interpretations of the underlying physics of cosmology and structure formation, as there are still many challenges that need to be addressed.

In particular, while we have assumed a fixed cosmology in this paper, a full analysis will need to model the spatial clustering of massive clusters as a function of cosmology. It should be noted that various simulations suites do now exist e.g., Aemulus (DeRose et al. 2019; McClintock et al. 2019c; Zhai et al. 2019; McClintock et al. 2019a), \textsc{Quijote} (Villaescusa-Navarro et al. 2020), Abacus (Garrison et al. 2021; Maksimova et al. 2021) to facilitate such modeling.

Obtaining accurate and unbiased halo mass estimates from the data will be another issue that needs to be addressed. One could use cluster scaling relations to make halo mass estimates from observables (e.g., richness \(\lambda\) in the case of redMaPPer, X-ray luminosity, gas mass, or temperature for X-ray observations Vikhlinin et al. 2006, 2009; Kim & Fabiano 2013; Kim et al. 2019, and Compton \(Y\) for SZ samples Planck Collaboration et al. 2014a; de Haan et al. 2016; Planck Collaboration et al. 2016b) which have different levels of systematics (see Giodini et al. 2013 for a review).

Conversely, one could also use galaxy–halo connection models such as halo-occupation-distribution models (HODs) or semi-analytic-models (SAMs) to map observables onto halos, but specific assumptions that go into the galaxy–halo connection model can also lead to various levels systematics (see Somerville & Davé 2015; Wechsler & Tinker 2018 for reviews). In addition, selection functions are not always well-studied as in the redMaPPer case, so applying them to the simulation is not always straightforward and adds an extra barrier to making apples-to-apples comparison between observations and simulations. Thus, it is promising yet challenging to make explicit cosmological constraints using \(k\)-NN CDFs in future work.

The main focus of this work is to detect auto-correlation clustering using \(k\)-NN summary statistics for the selected redMaPPer clusters. In addition to the query-data point distances used in this work, a recent development n the \(k\)-NN formalism replaces the query points with a second data set and allows cross-correlations to be measured between two clustered samples (Banerjee & Abel 2021b). An example appli-
cation where the kNN cross-correlation formalism could potentially improve over the two-point correlation function (Li et al. 2019; Rafiei-Ravandi et al. 2021) is in the localization of fast-radio-bursts (FRBs) (sparse data) to their host galaxies in a galaxy survey (dense data). Thus, kNN summary statistics is almost certain to have a broad impact on improving our understanding of cosmology given its myriad of merits as discussed in this work.

ACKNOWLEDGEMENTS

We thank Adam Mantz, Chun-hao To, Daniel Gruen, Si-han (Sandy) Yuan, Steven Allen, and Yuuki Omori for helpful discussions and comments while preparing this draft. We thank the anonymous referee for a careful read of the manuscript and insightful comments that helped to improve the draft. YW acknowledges the support of a Stanford-KIPAC Chabolla Fellowship. This work was supported by the Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, the U.S. Department of Energy (DOE) Office of Science Distinguished Scientist Fellow Program, and the U.S. Department of Energy SLAC Contract No. DE-AC02-76SF00515. This research made use of the Sherlock cluster at the Stanford Research Computing Center (SRCC); the authors are thankful for the support of the SRCC computing team. This research made extensive use of arXiv.org and NASA’s Astrophysics Data System for bibliographic information.

DATA AVAILABILITY

Data presented in this work will be available upon reasonable request with the authors. Alternatively, readers are welcome to reproduce the results shown in this work with our code and the original data of redMaPPer.  

REFERENCES

Abbott T. M. C., et al., 2018, Phys. Rev. D, 98, 043526
Alhara H., et al., 2011, ApJS, 193, 29
Allen S. W., Evrard A. E., Mantz A. B., 2011, ARA&A, 49, 409
Banerjee A., Abel T., 2021a, MNRAS, 500, 5479
Banerjee A., Abel T., 2021b, MNRAS, 504, 2911
Banerjee A., Kokron N., Abel T., 2021, arXiv e-prints, p. arXiv:2107.10287
Bleem L. E., et al., 2015, ApJS, 216, 27
DeRose J., et al., 2019, ApJ, 875, 69
Driver S. P., et al., 2011, MNRAS, 413, 971
Eisenstein D. J., et al., 2011, AJ, 142, 72
Euclid Collaboration et al., 2019, A&A, 627, A23
Garrison L. H., Eisenstein D. J., Ferrer D., Maksimova N. A., Pinto P. A., 2021, MNRAS, 508, 575
Gaztaña E., Cabré A., Hui L., 2009, MNRAS, 399, 1663
Giodini S., Lovisari L., Pointecouteau E., Ettori S., Reiprich T. H., Hoekstra H., 2013, Space Sci. Rev., 177, 247
Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399
Hilton M., et al., 2018, ApJS, 235, 20
Hilton M., et al., 2021, ApJS, 253, 3

6 https://github.com/Bulk826R/halo_vpf
7 http://risa.stanford.edu/redmapper/
APPENDIX A: K-MEANS CLUSTERING
JACKKNIFE METHOD
We quoted Jackknife errors for the $\xi(r)$ measurements in Section 4.2 as an estimate of the intrinsic scatter in the 1000 richest redMaPPer clusters. We use the k-means clustering method for obtaining 200 Jackknife samples of the redMaPPer clusters by dividing the 1000 richest redMaPPer clusters into 200 patches based on their RA and DEC (Fig. A1 upper panel). Each of the 200 Jackknife sub-samples are constructed by leaving one patch of clusters out, forming two hundred 199-patch sub-samples in total. Measuring $\xi(r)$ for all 200 Jackknife sub-samples will create the Jackknife errors in each radial bin. We deliberately chose 200 patches to avoid large angular size patches created (when patch number < 100) that preferentially remove the more abundant low-mass clusters and bias the Jackknife sub-samples’ $\xi(r)$ above the clustering signal of the redMaPPer clusters. The average patch size seen in the top panel of Fig. A1 is $\sim 10^\circ$ and converts to a physical size of $\sim 150$ Mpc for individual patches at $z = 0.2$ (median redshift for the selected redMaPPer clusters is $z = 0.23$), which is comparable to the largest scales probed by $\xi(r)$ in Fig. 5.

APPENDIX B: COMPARISON WITH SIMULATIONS
In this appendix section we describe an ancillary effort to measure the kNN CDFs and $\xi(r)$ for a simulated cluster sample as a test of our methodology. We select simulated halos from the dark-matter-only cosmological simulation suite QUIJOTE (Villaescusa-Navarro et al. 2020). This is a set of 1 (Gpc/h)$^3$-box simulations with varying cosmological parameters. Since our analysis of the redMaPPer clusters are conducted under fixed cosmology, we select simulated halos from a set of simulation runs at a fixed fiducial flat ΛCDM cosmology (cosmological parameter values: $\Omega_m = 0.3175$, $h = 0.6711$, $\sigma_8 = 0.834$, $n_s = 0.9624$). The number density of the simulated halos are chosen to match that of the 1000 richest redMaPPer clusters in the survey footprint as described in Section 3.1, resulting in selecting the 526 most massive clusters from the simulations. This cluster sample has a minimum halo mass of $4.6 \times 10^{14} M_\odot/h$ (Friends-of-Friends group mass).

For the kNN CDFs (Fig. B1) and the two-point correlation function (Fig B2), we choose the same radial ranges as we did for the redMaPPer clusters in Section 4 for both summary statistics over which we evaluate the clustering. We use 526 random samples with 526 random points each to measure the non-clustered result for both kNN CDFs and $\xi(r)$. All calculations for kNN CDFs and two-point correlation functions assume non-periodic boundary conditions to maximally mimic the signal extraction procedure in redMaPPer. Any additional differences with redMaPPer may arise from survey volume geometry, selection masks, mass/number density limit adopted for the simulated clusters, the underlying cosmology assumed for the simulations etc.

For the two-point correlation function (Fig. B3), the intrinsic scatter (obtained by sampling the 526 most massive clusters from 526 boxes run with the same cosmology but different initial conditions) of the 526 most massive clusters behave similar to the redMaPPer clusters, which is not significant (within 1σ errors compared to the randoms) throughout the radial range but shows a slight increase in the clustering strength towards small scales. This is consistent with the redMaPPer clusters being dominated by shot noise at scales of $r \gtrsim 70$ Mpc (Fig. 5).

As for the kNN CDFs, the clustering signal of the 526 most massive simulated clusters demonstrate similar features as that of the redMaPPer clusters (Fig. 4) for both 1NN and 2NN relative to the randoms. We note that the signal of the simulated clusters are different to the observed ones mainly owing to the differences in the volume geometry occupied by the clusters (periodic cubic box versus non-periodic conic-shaped survey footprint). Also, we explore two sources of intrinsic scatters for the kNN CDFs. The shaded regions in Fig B1 denote the 1σ Jackknife errors of the simulated clusters’ kNN CDFs by leaving out the $k = 1$ or 2 NN distances associated with one particular cluster at a time (method same as in Section 4.1, but since we have only 526 simulated clusters, there are only 526 leave-one-out Jackknife samples instead of 1000 in the redMaPPer case). The solid error bars denote the 1σ scatter for the kNN CDFs of the 526 most massive simulated clusters due to different initial random seeds of the simulation under the same cosmology (526 different realizations). As shown in the figure, the Jackknife error from the randomization of query points is slightly larger than the scatter caused by varying the initial conditions of the clusters on small scales ($r \lesssim 70$ Mpc), and vice versa for larger scales in both the $k = 1$ and 2 cases. Therefore, the leave-one-out Jackknife error estimate for the redMaPPer clustering signals shown in Fig. 4 well-represent the intrinsic scatter under fixed cosmology.

We also explore the non-Gaussian clustering signal detected in the redMaPPer clusters in Section 5 in the QUIJOTE simulations. We choose the same radial range for the QUIJOTE boxes as the redMaPPer clusters, although we switch from the leave-one-out Jackknife method to the simulation-to-simulation intrinsic scatter for a more realistic estimation

\[^8\] https://github.com/esheldon/kmeans_radec
Figure B1. The relative CDFs for the simulated cluster/random points versus the average CDF for the random points. Results for the 1NN and 2NN CDFs are shown in the left and right panels, respectively (identical radial ranges as in Fig. 4). In each panel, the thick solid line shows the relative CDF of the 526 most massive simulated clusters to the mean of the 526 random samples. The colored shaded region denotes the 1σ Jackknife errors obtained by leaving out query points associated with one cluster at a time. The solid error bars denote the 1σ simulation-to-simulation (Sim-to-sim in legend) scatter of 526 realizations of the 526 simulated clusters with different initial conditions under the same cosmology (fixed query points), which is nearly identical to the leave-one-out Jackknife errors. The shape of the clustering signal, as well as the magnitude of the Jackknife errors, are comparable to the redMaPPer clusters CDFs shown in Fig. 4.

APPENDIX C: TWO-POINT CORRELATION FUNCTION AT SMALL SCALES

As shown in Section 4.2, the two-point correlation function for the 1000 richest redMaPPer clusters show a slight increase of the clustering signal towards smaller scales (top panel of Fig. 5). A similar trend for the number-density-matched simulated clusters was also observed (Fig. C1). Therefore, we investigate in this appendix section how the two-point correlation function signal responds to different scale cuts at small scales.

In Fig. C1, we show the two-point correlation function (upper panel) and the $\chi^2$ distribution (lower panel) of the 1000 richest redMaPPer clusters compared with the 2000 random samples, which is obtained by extending the lower limit of the radial scale down to 20 Mpc. We keep the same radial

scale upper limit (155 Mpc) and number of bins (50) to control the variation of other scale-related factors. We find that the clustering signal for the redMaPPer clusters continue to become more significant below $r = 35$ Mpc, and that their $\chi^2$ of $\xi(r)$ increases from $\chi^2 \sim 110$ (lower panel of Fig. 7) to $\chi^2 \sim 250$ (lower panel of Fig. C1). Although this clustering signal is still less significant than what was detected by the joint $k = 1, 2$-NN CDFs (upper panel of Fig. 7), it elaborates that the two-point correlation function is more sensitive to clustering at the small scales ($r \lesssim 50$ Mpc), where most of the clustering signal is dominated by close cluster pairs in high density regions. We have also verified that the $k$NN CDFs’ $\chi^2$ improves only marginally with the same small scale extension, which is due to the $k = 1, 2$-NN CDFs we used being quite insensitive to overdensities than higher order $k$s. This suggests that the low-$k$ nearest neighbor CDFs complements the two-point correlation function at large scales whereas $\xi(r)$ is more sensitive to clustering at small scales.

APPENDIX D: VISUALIZATION OF $k$NN DISTANCES

We present a visualization of $k$NN distances measured from a set of query points located at fixed redshift. In this example, we select the RA and DEC of all healpix anchors that belong to the SDSS footprint and satisfy our selection criteria in Section 3.3 (NSIDE=2048, $f_{\text{good}} > 0.5$, $s_{\text{max}} \geq 0.3$). We assume that these query points are located at a fixed redshift of $z = 0.2$ (distance of 842.87 Mpc). This results in 12779404 query points at $z = 0.2$, from which we measure the $k = 1, 2, 3$-NN distances to the 1000 richest redMaPPer clusters. The results are shown in Fig. D1.
The four panels show the color maps of the 1NN, 2NN, 2NN-1NN (‘-’ for minus), and 3NN-2NN distances, respectively. For example in the top left panel, the small 1NN distances (orange shade) mark out regions close to clusters and stand for overdensities, whereas large 1NN distances (purple shade) mark out underdense voids. Similar behavior is observed for the 2NN distances, although the voids marked out by 2NN is more significant than 1NN. As for the kNN distance differences (2NN-1NN and 3NN-2NN), the color maps show the contrasting features of different ks at each query point, and this should not be misinterpreted as the CIC functions $P_{k|V}$. For example, if a query point lives near a cluster, its 1NN distance will be close to 0 whereas its 2NN distance will approach the mean inter-cluster distance, leading to large 2NN-1NN values at the locations of overdensities seen in the top row. The query points located in underdense regions will also automatically trace out the Voronoi tessellation of the clusters, as query points sitting on Voronoi cell walls between

**Figure B2.** Upper panel: The measured $k = 1, 2, 3, 4$-NN CDFs for the 526 most massive clusters in the Quijote simulation (one box) is denoted by the solid curves. The shaded regions around them indicate the 1σ sim-to-sim scatter from 526 simulation boxes under fixed cosmology but with different initial conditions. The dotted turquoise and green curves denote the predicted 3NN(1,2) and 4NN(1,2) CDFs assuming a random Gaussian field traced by the simulated clusters, which are derived from the measured 1NN and 2NN CDFs propagated through equations 12 and 13. Lower panel: The $\chi^2$ distribution of the joint $k = 3, 4$-NN CDFs. The blue histograms denote the $\chi^2$ of 526 Jackknife it measurements of the 3NN and 4NN CDFs, while the dashed red line marks the $\chi^2$ for the Gaussian field-prediction of the 3NN and 4NN CDFs. The predicted CDFs deviate significantly from the measurements and corroborates the detection of non-Gaussian clustering as in redMaPPer.

**Figure B3.** Upper panel: The two-point correlation function measurements of the 526 most massive clusters (matching the selected redMaPPer clusters’ number density) in Quijote simulation (blue dotted-dashed curve) versus 526 samples of random points (red, 526 each). The shaded region around the random’s $\xi(r)$ show the 1σ distribution of all samples while the dashed curve denotes the mean. The error bar around the simulated clusters denote the intrinsic scatter ($1\sigma$) of the signal, which is obtained by sampling the 526 most massive clusters from 526 simulation boxes run with the same cosmology but different initial conditions. Lower panel: The deviation of the $\xi(r)$ signal along with its intrinsic scatter for the 526 most massive simulated clusters in units of the randoms’ standard deviation as a function of radial scale. Large scales ($r \gtrsim 70$ Mpc) of the simulated clusters’ $\xi(r)$ is submerged by shot noise due to limited sample size.
Figure C1. Two point correlation function measurement of the 1000 richest redMaPPer clusters down to $r = 20$ Mpc. Upper panel: Same as the top panel of Fig. 5, but now the lower limit of the radial range is extended down to 20 Mpc. Lower panel: The $\chi^2$ distribution of the two-point correlation of the random mock clusters, versus that of the redMaPPer clusters, both measured down to 20 Mpc. Compared to the lower panel of Fig. 7, the signal significance ($\chi^2$) of the redMaPPer clusters’ $\xi(r)$ increased by $\sim 140$ due to the extension of the two-point function to smaller radial scales.

two clusters will have equal 1NN and 2NN distances. The 3NN-2NN distances demonstrate a more complicated tessellation of the clusters as traced by the query points.

This paper has been typeset from a \TeX/\LaTeX file prepared by the author.
Figure D1. The $k$NN distances to the 1000 richest redMaPPer clusters measured from $1.3 \times 10^8$ query points located at $z = 0.2$ (842.87 Mpc). The query points are all healpy pixels of the SDSS survey footprint with $f_{\text{good}} > 0.5$ and $z_{\text{max}} \geq 0.3$. The color bars denote the $k$NN distances plotted in each panel. **Top left:** 1NN distance. **Top right:** 2NN distance. **Bottom left:** The difference of the distance to the 2NN and 1NN clusters for every query point. **Bottom right:** The difference of the distance to the 3NN and 2NN clusters for every query point.