# A Widely Applicable Galaxy Group Finder Using Machine Learning: Application to SDSS DR13 Postprint

**Authors:** Juntao Ma, Jie Wang, Tianxiang Mao, Hongxiang Chen, Yuxi Meng, Cheng Li, Xiaohu Yang and Qingyang Li

**Date:** 2025-06-13T16:51:13+00:00

## Abstract

We present the application of a machine learning based galaxy group finder to real observational data from the Sloan Digital Sky Survey Data Release 13 (SDSS DR13). Originally designed and validated using simulated galaxy surveys in redshift space, our method utilizes deep neural networks to recognize galaxy groups and assess their respective halo masses. The model comprises three components: a central galaxy identifier, a group mass estimator, and an iterative group finder. Using mock catalogs from the Millennium Simulation, our model attains above 90% completeness and purity for groups covering a wide range of halo masses from 1011 to 1015 h−1M . When applied to SDSS DR13, it successfully identifies over 420,000 galaxy groups, displaying a strong agreement in group abundance, redshift distribution, and halo mass distribution with conventional techniques. The precision in identifying member galaxies is also notably high, with more than 80% of groups with lower mass achieving perfect alignments. The model shows strong performance across different magnitude thresholds, making retraining unnecessary. These results confirm the efficiency and adaptability of our methodology, offering a scalable and accurate solution for upcoming large-scale galaxy surveys and studies of cosmological formations. Our SDSS group catalog and the essential observable properties of galaxies are available at https://github.com/JuntaoMa/SDSS-DR13-group-catalog.git.

## Full Text

## Abstract

We present the application of a machine learning based galaxy group finder to real observational data from the Sloan Digital Sky Survey Data Release 13 (SDSS DR13). Originally designed and validated using simulated galaxy surveys

in redshift space, our method utilizes deep neural networks to recognize galaxy groups and assess their respective halo masses. The model comprises three components: a central galaxy identifier, a group mass estimator, and an iterative group finder. Using mock catalogs from the Millennium Simulation, our model attains completeness and purity above 90% for groups covering a wide range of halo masses from $10^{11}$ to $10^{15}$ $h^{-1}$Me.

When applied to SDSS DR13, it successfully identifies over 420,000 galaxy groups, displaying strong agreement in group abundance, redshift distribution, and halo mass distribution with conventional techniques. The precision in identifying member galaxies is also notably high, with more than 80% of lower mass groups achieving perfect alignment. The model shows strong performance across different magnitude thresholds, making retraining unnecessary. These results confirm the efficiency and adaptability of our methodology, offering a scalable and accurate solution for upcoming large-scale galaxy surveys and studies of cosmological formations. Our SDSS group catalog and the essential observable properties of galaxies are available at https://github.com/JuntaoMa/SDSS-DR13-group-catalog.git.

Key words: (cosmology:) large-scale structure of universe – Galaxy: halo – methods: data analysis

## 1. Introduction

Current models of cosmic structure formation depict galaxies not as isolated entities but as integral components of larger gravitationally bound systems, known as galaxy groups. Galaxy groups typically comprise a few to dozens of galaxies residing within a dark matter halo, playing an essential role in galaxy evolution through interactions such as mergers, tidal forces, and ram pressure stripping. Understanding galaxy groups is therefore critical for elucidating galaxy formation processes, environmental influences on galaxy properties, and the broader context of the cosmic web, where groups represent key structural elements linking individual galaxies to clusters and large-scale filaments.

Over past decades, researchers have developed numerous algorithms for galaxy group identification in large redshift surveys. Early pioneering works introduced the Friends-of-Friends (FoF) algorithm (e.g., Huchra & Geller 1982; Eke et al. 2004; Knobel et al. 2009), a percolation-based method widely applied due to its simplicity and effectiveness. In FoF, galaxies within a predefined linking length are iteratively grouped, creating clusters based purely on spatial proximity. Halo-based group finders, such as the algorithm developed by Yang et al. (2005), utilize physical models of dark matter halos (e.g., the Navarro–Frenk–White profile) to iteratively assign galaxies to host halos according to estimated halo mass and galaxy distributions. These classical methods have successfully generated group catalogs from numerous redshift surveys, including CfA redshift survey (Huchra & Geller 1982), the Two Degree Field Galaxy Redshift Survey (2dFGRS; e.g., Eke et al. 2004; Yang et al. 2005; Tago et al. 2006), the

Two Micron All Sky Redshift Survey (2MRS; e.g., Crook et al. 2007; Lavaux & Hudson 2011; Tully 2015), the Sloan Digital Sky Survey (SDSS; e.g., Goto 2005; Miller et al. 2005; Berlind et al. 2006; Yang et al. 2007; Lim et al. 2017), the zCOSMOS (Wang et al. 2020) and the DESI Legacy Imaging Surveys (Yang et al. 2021).

Despite their widespread use, classical group-finding methods have notable limitations. Primarily, these algorithms are sensitive to the choice of parameters, such as linking lengths or halo mass-to-light ratios, and often require extensive calibration against simulations. To address these challenges, recent studies have begun incorporating machine learning techniques, particularly artificial neural networks (ANNs), into galaxy group identification. ANNs can effectively learn intricate patterns directly from observational data, reducing reliance on manually tuned parameters. Various network architectures—including multilayer perceptrons, convolutional neural networks (CNNs; Arbib 1995), graph neural networks (GNNs; Bronstein et al. 2017), and recurrent neural networks (RNNs; Hochreiter & Schmidhuber 1997)—have demonstrated robust performance across diverse astrophysical applications.

CNNs have shown significant success in object detection tasks in astronomical imaging, while graph neural networks effectively handle spatially related data. For example, Mao et al. (2021) introduced an innovative convolutional neural network framework for reconstructing baryon acoustic oscillation (BAO) signals, significantly enhancing the BAO signal-to-noise ratio to around k  0.4 h $Mpc^{-1}$. Chen et al. (2024) applied ANNs to evaluate the environmental characteristics of galaxies. This approach allowed for precise estimation of line-of-sight velocities and facilitated the reconstruction of the real-space power spectrum with an error less than 5% at s $>$ 8 $h^{-1}$ Mpc.

In our previous study, we presented a galaxy group identification tool based on ANNs, which we trained and substantiated using detailed cosmological simulations (Ma et al. 2025; referred to hereafter as Paper I). The inherent nonlinear capabilities of ANNs enabled our algorithm to discern intricate patterns within the data set. Our assessment on simulation data revealed the model's exceptional precision in assigning member galaxies and estimating halo masses, accurately identifying over 92% of galaxy groups and achieving halo mass errors of less than 0.3 dex. Importantly, the machine learning based method demonstrated significant versatility, requiring minimal recalibration across various data sets. Without the need for additional training, we verified its competence in producing dependable results on sparse samples with brighter magnitude cut up to m_r $<$ 14, high-redshift samples up to z = 1.08, and data sets from different simulations.

These evaluations highlight the reliability and wide-ranging applicability of our group finder. Furthermore, the group finder performs well on redshift-distorted samples when re-trained on the corresponding data sets.

In this paper, we extend the work presented in Paper I by applying our group

finder to redshift-space data and testing its performance on real observations from the Sloan Digital Sky Survey Data Release 13 (SDSS DR13; Albareti et al. 2017). SDSS DR13 offers extensive spectroscopic data, providing an ideal benchmark for evaluating the effectiveness of our method in practical applications. Using the galaxy catalog of SDSS DR13 provided by Lim et al. (2017), we assess the algorithm's ability to reliably identify galaxy groups in a real survey environment. The structure of this paper is outlined as follows: Section 2 provides an overview of both the simulation data set employed for model training and the SDSS data set. Section 3 describes the architecture of our machine learning based group finder and details of the procedures. In Section 4, we assess the effectiveness of our method using mock galaxy catalogs from simulations, focusing on aspects like completeness, purity, and halo-mass assignment. Section 5 presents the outcomes of our group catalog and compares these results to conventional techniques. Lastly, Section 6 discusses our main findings and suggests potential avenues for future research.

## 2. Data

This section presents the data sets employed in our study, especially the mock galaxy catalogs used for training and testing from the Millennium simulation. Also, the data from the real survey Sloan Digital Sky Survey (SDSS) is described in detail as follows.

### 2.1. Simulation and Mock Galaxy Catalog

The simulation used in this work is Millennium Simulation (MS; Springel et al. 2005), which is a large scale N-body simulation of cosmic structure formation based on ΛCDM cosmology. The simulation consists of $N = 2160^3$ dark matter particles with particle mass $8.6 \times 10^8$ $h^{-1}$Me, in a comoving volume of $(500$ $Mpc/h)^3$. These particles evolve from $z = 127$ to $z = 0$. The cosmological parameters of the simulation adopted by MS are the following: $\Omega\_m = 0.25$, $\Omega\_b = 0.045$, $h = 0.73$, $\Omega\_\Lambda = 0.75$, $n = 1$, and $\sigma\_8 = 0.9$, and the Hubble constant is defined as $H_0 = 100$ h km $s^{-1}$ $Mpc^{-1}$.

We use the semi-analytic galaxy catalog of MS developed by Guo et al. (2011), which implements the galaxy formation model L-Galaxies (Henriques et al. 2015) onto merger trees extracted from the MS. To train and evaluate our machine-learning-based galaxy group finder, we constructed mock galaxy catalogs derived from the MS. Compared to the redshift-distorted samples used in Paper I, which were directly extracted from a sub-box of the simulation, the mock catalogs in this work are generated using a different approach to more realistically reproduce the observational characteristics of SDSS. We stack the simulation box into a larger cubic volume (33 times the original volume) to achieve sufficient survey depth out to a redshift limit of $z = 0.2$. We selected only groups with host halos containing more than 100 dark matter particles and also apply an r-band absolute magnitude cut of $M\_r < -14$ to exclude galaxies that may be unreliable in simulations. The observer is placed at the center of this enlarged

simulation volume. We incorporate redshift-space distortions by adjusting each galaxy's cosmological redshift ($z_0$) according to its line-of-sight peculiar velocity (v_r): z_{rsd} = $z_0$ + v_r/c. A redshift cut of $z \leq 0.2$ is applied to match the observational depth of SDSS.

Galaxy apparent magnitudes were computed using their absolute magnitudes and luminosity distances, adopting r-band magnitude limit of m_r < 17.7, consistent with the SDSS observational selection criteria (Abazajian et al. 2009). To robustly assess the generalization capability of our model, we select five distinct sky areas within the constructed mock survey. The largest of these, covering approximately 12,000 square degrees, forms our primary training data set and contains a total of 875,520 galaxies. The other four areas, each spanning roughly 7000 square degrees, serve as independent test data sets containing 401,027; 407,432; 394,189; and 417,336 galaxies, respectively.

### 2.2. SDSS Data

The redshift survey data set employed in this study is drawn from the SDSS DR13, which represents the first data release of the fourth phase of the Sloan Digital Sky Survey (SDSS-IV). The data set provides photometric observations in five broad optical bands (u, g, r, i, z). In DR13, photometric calibration was improved through updated zero-points and flat fields, as implemented via the hypercalibration procedure described by Finkbeiner et al. (2016). Additionally, DR13 includes redshift measurements for galaxies previously excluded from spectroscopic observations in DR7 due to fiber collisions—i.e., cases where galaxies had close neighbors within the minimum fiber separation.

In order to validate our group-finding method, we utilize galaxies of the SDSS sample in Lim et al. (2017) and compare with their group catalog. This sample, derived from the Legacy Survey region in DR13, encompasses roughly 23% of the sky with an r-band magnitude limit set at 17.77. Some galaxies in this sample lack spectroscopic redshift measurements due to various issues such as fiber collisions, broken or unplugged fibers, low-quality spectra, or poor model fits. For these galaxies, redshifts from external sources were assigned, with the sample restricted to $z \leq 0.2$. Consequently, there are two catalogs in Lim et al. (2017): the SDSS sample, containing only galaxies with spectroscopic redshifts obtained from SDSS DR13 or other sources, and the SDSS+ sample, which includes all galaxies, supplemented with redshifts estimated from nearest neighbors and from 2MASS Photometric Redshift catalog (2MPZ; Bilicki et al. 2014). Throughout this work, we utilize only the SDSS sample from Lim et al. (2017).

In our schemes, stellar mass of galaxy is utilized as an input instead of luminosity. It is important to emphasize that the stellar mass in the SDSS sample is derived using an empirical relationship detailed by Bell et al. (2003): log(M_*/M_ ) = -0.499 + 1.519(g-r) - 0.586$(g-r)^2$ - 0.5, where 4.67 is the absolute magnitude of the Sun in the r band.

## 3. Method

The framework of our group finder consists of two machine-learning models: (1) Central Galaxy Identifier, predicting the central galaxies of all galaxies in the catalog; (2) Group Mass Estimator, predicting the halo mass of identified groups. Our final group catalog is based on the prediction of these two networks, by applying an iteration process to merge the potential groups identified by the network. This framework is the same as what we did in Paper I, only with some minor modifications in the Group Mass Estimator. In addition, the model is trained on mock data sets specifically designed to better replicate the properties of SDSS observations, as described in Section 2.1. We briefly show the algorithm of our schemes and the modification as follows.

### 3.1. Central Galaxy Identifier

A group catalog can be complex given the large number of galaxies, but the groups can be expressed as center-satellite pairs of galaxies. Therefore, we build a network to identify the central galaxies of all galaxies from their nearest neighbors. Specifically, we search for neighbor galaxies within a line-of-sight distance of $\pm10h^{-1}$ Mpc from each target galaxy. These galaxies are ranked by their projection distances to the target galaxy. The properties of the nearest ten neighbors and the target galaxy are taken as the network's input features. We list all the inputs as follows: r-band magnitude of target galaxy; color of target galaxy; redshift of target galaxy; projection distance to its i-th neighbor; r-band magnitude of i-th neighbor; line of sight distance to its i-th neighbor; color of its i-th neighbor.

The ten neighbor galaxies of the target galaxy are numbered from 1 to 10, according to ranks of projection distances in ascending order. The target galaxy itself is indexed as 0. If the central galaxy is among these 11 galaxies and its index is i, we will categorize the target galaxy as i-th class. In the situation that none of these galaxies are its center, the target galaxy will be assigned to class 11. The architecture of our neural network incorporates four hidden layers, each utilizing Rectified Linear Unit (ReLU) activation functions. The output layer generates a 12-dimensional probability vector corresponding to each class. We implement cross-entropy loss for optimization and train the model over 500 epochs using the training data set from the MS.

### 3.2. Group Mass Estimator

The host halo mass of galaxies is a fundamental quantity for understanding galaxy formation, evolution, and environmental effects. To accurately estimate the halo mass of galaxy groups identified from observational data, we have developed an ANN aimed at halo mass prediction. In this work, we use the term M_{vir} to refer to the halo mass defined as M_{200}, i.e., the mass enclosed within a spherical region whose average density is 200 times the critical density. This neural network leverages key observational properties of galaxy groups,

particularly focusing on the characteristics of the central galaxy and its largest satellite galaxies. In our group-finding procedure, a group is defined as a set of galaxies assigned to the same central galaxy by the Central Galaxy Identifier. The inputs to the network consist of the properties of the central galaxy and the N most massive satellite galaxies within the group: stellar mass of central galaxy; total stellar mass of all galaxies in the group; total number of galaxies in the group; projection distance of i-th satellite to group center; redshift of the central galaxy; stellar mass of i-th satellite galaxy.

Here, the index i corresponds to the rank of the satellite galaxy by stellar mass, with i = 1 representing the most massive satellite. Groups with higher richness, i.e., more member galaxies, benefit from using a larger N; however, the prediction accuracy for low-richness groups may degrade if N is too large. We adopt N = 5 as a compromise to balance performance across groups of varying richness. For groups containing fewer than five satellite galaxies, the missing values of M_*,i and d_i are set to zero.

Compared to our previous work, we have refined the input parameters of our mass prediction model. Specifically, we replaced the maximum r-band absolute magnitude (M_r,max) among group galaxies with the redshift of the central galaxy. Additionally, when selecting the most massive satellite galaxies, we exclude the outermost 50% (rounded down) of satellites based on their projected distances to center. This adjustment mitigates potential inaccuracies in halo mass estimation arising from uncertainties in the measurements of input properties. For instance, in the SDSS data set, a small number of galaxies are assigned extremely faint magnitudes (e.g., M_r > -10), which can confuse the original model in Paper I and lead to unreliable halo mass predictions. The performance of both the original and updated mass models will be presented in the following section.

The neural network architecture comprises four hidden layers, each utilizing ReLU activation functions to introduce nonlinearity. We employ the Mean Squared Error as the loss function during training. The ANN is trained on the training data set derived from the MS for 500 epochs, ensuring robust generalization and reliable halo mass predictions for groups identified in observational catalogs such as SDSS.

[Figure 1: see original paper]. Global completeness and purity across four mock test catalogs. Group completeness and purity are defined in the main text. Dashed lines represent the mean values, and error bars denote the $1\sigma$ standard deviation among the four data sets. Completeness (blue) is plotted as a function of the true virial mass of groups, M_{vir},t, while purity (green) is shown as a function of the predicted virial mass of identified groups, M_{vir},p. All groups achieve completeness and purity levels exceeding 90% across the full mass range.

### 3.3. Identification of Groups

The Central Galaxy Identifier model provides central-satellite relationship data, enabling highly precise identification of galaxy groups. Nonetheless, certain satellites are somewhat distant from their centers, causing these centers to fall outside the top ten nearest neighbors. Consequently, this could lead to the fragmentation of the group into multiple smaller subsets. To mitigate this limitation, we have implemented a simplified iterative merging process, outlined as follows:

1. **Initial Estimation**. The groups predicted by the Central Galaxy Identifier are considered as candidate groups. Use the Group Mass Estimator to predict halo masses (M_{vir}) and corresponding radii (R_{vir}) for each candidate group.
2. **Group Merging**. Combine groups that overlap spatially within their R_{vir} radii with line-of-sight tolerances of $\pm 5h^{-1}$ Mpc, merging them into a single group centered on the one with the highest estimated halo mass.
3. **Iterative Refinement**. Recalculate halo masses for the merged groups, repeating the merging process until no further changes occur, resulting in a stable final catalog.

## 4. Performance on Mock Catalog

To evaluate the performance of our group finder, we adopt the methodology outlined in Wang et al. (2020). First, we define several notations for clarity:

1. **IG**. An identified galaxy group produced by the group finder.
2. **TG**. A true galaxy group from the simulation, corresponding to a known host halo.
3. **IG-T**. An identified group successfully matched to a true group.
4. **TG-I**. A true group successfully matched to an identified group.

Matching IGs and TGs is nontrivial because groups often differ slightly in membership composition. We utilize a combined approach involving two matching criteria: member matching and central matching. Member matching requires that more than 50% of the members in an IG are also members of a TG, and vice versa. Central matching demands that the central galaxy of an IG is the same as that of a TG. While member matching is typically more stringent and reliable, especially for large groups, combining it with central matching helps enhance overall robustness. In cases where the two criteria yield differing results, priority is given to member matching outcomes. Through this combined matching strategy, we establish clear one-to-one pairings between identified and true groups, referred to as IG-T and TG-I pairs, respectively.

Our group finder is analyzed by employing four distinct mock catalogs derived from simulations to assess its performance. Initially, we focus on the global com-

pleteness and purity of the galaxy groups that our method identifies. Global completeness refers to the ratio of true groups (TGs) in the simulations that our method successfully matches to identified groups (IGs). On the other hand, global purity indicates the proportion of IGs that accurately correspond to the true groups in the simulations. In Figure 1, we depict the global completeness and purity as functions of halo mass. The graph presents the average completeness and purity for all four test data sets using dashed lines, with error bars denoting the standard deviation. Notably, both completeness and purity are consistently above 90% across the entire spectrum of halo mass ranges considered, even for groups with masses as low as M_{vir}  $10^{11}$ $h^{-1}$Me. Moreover, we notice a rising trend in these metrics for increasing halo masses, underscoring the dependability of our identification method for larger groups. This observation is particularly significant, given that massive groups are critical in mapping the cosmic density field.

Next, we further examine the accuracy of our identified groups by analyzing individual galaxy memberships. For each IG matched with a corresponding TG, we define two membership accuracy metrics. Consider an identified group containing N_i predicted member galaxies matched to a true group containing N_t galaxies. If N_s galaxies exist both in the identified and true groups, we define:

1. **Member Completeness**. f_c = N_s / N_t
2. **Member Purity**. f_p = N_s / N_i

[Figure 2: see original paper]. Cumulative distributions of member completeness and purity across four mock test catalogs. Left panel: The x-axis represents member completeness (f_c), while the y-axis shows the fraction of groups with completeness greater than a given value f_c. Results are displayed for four halo mass bins, distinguished by different colors. Solid lines denote the mean values and error bars indicate the $1\sigma$ standard deviation across the four test catalogs. Right panel: Similar to the left panel, but illustrating the cumulative distribution of member purity (f_p) for the same halo mass bins.

In Figure 2, we present the cumulative distributions for member completeness (f_c) and purity (f_p), categorized into four halo mass ranges, each represented by distinct colors. Our group finder demonstrates high precision in assigning member galaxies, particularly for groups with masses below $10^{14}$ $h^{-1}$Me. In excess of 80% of these groups achieve perfect membership accuracy, where both completeness (f_c) and purity (f_p) reach 1. As we examine groups with higher masses, the task of member galaxy assignment becomes more challenging due to an increased number of satellites and amplified effects of redshift-space distortions. This complexity reduces the percentage of groups with perfect completeness and purity to about 40%. Nevertheless, even under these demanding conditions, over 80% of the groups uphold complete and pure values above 0.6, highlighting the consistent effectiveness of our group finder across an extensive range of group masses.

Figure 3 presents a direct comparison between the true and predicted halo masses of matched groups, along with the results from the redshift-distorted samples in Paper I. In this work, the comparison is shown as a 2D histogram plot, while the relation from Paper I is represented by median values with error bars. The updated model demonstrates performance comparable to that of the original version in terms of mass prediction in mock catalog, which is expected.

[Figure 3: see original paper]. Comparison of true and predicted group masses across four mock test catalogs, along with results from Paper I. The upper panel shows a 2D histogram comparing the true and predicted halo masses of matched groups in this work. For reference, results from the redshift-distorted samples used in Paper I are overlaid as red points with error bars, representing median halo mass predictions with asymmetric error bars estimated from the 25th and 75th percentiles. The lower panel presents the standard deviation of log(M_{vir},p/M_{vir},t) for both data sets, quantifying the scatter in mass predictions.

Figure 4 presents a comparison between the distribution of predicted halo mass (M_{vir},p) and the actual halo mass obtained from simulations. Our model estimates halo masses solely based on the intrinsic properties of groups, without being trained on the true halo mass distribution, thus ensuring reliable predictions, especially in areas with limited observational data. This approach, however, can lead to some systematic deviations from the actual distribution. The figure demonstrates that although the predicted and actual distributions are quite similar, there is a slight underestimation for groups near M_{vir} $10^{14}$ $h^{-1}$Me. To potentially improve the accuracy and precision of our halo mass predictions, we could consider including additional variables, such as the environmental context of the target groups.

[Figure 4: see original paper]. Comparison between the predicted halo mass distribution and the true distribution across four mock test catalogs. The mass range [$10^{11}$ $h^{-1}$Me, $10^{15 \cdot 5}$ $h^{-1}$Me] is divided into 18 logarithmic bins. In the upper panel, the gray shaded region represents the $1\sigma$ variation in the true halo mass distribution across four test data sets. Blue points indicate the mean predicted distribution, with error bars showing the corresponding $1\sigma$ scatter among the test data sets. The lower panel shows the fractional difference between the two distributions, computed as the difference divided by the true halo mass distribution.

## 5. SDSS Group Catalog

In this section, we further discuss the group catalog that our group finder applies to the observational SDSS data set addressed in Section 2.2. Among 586,025 galaxies in the SDSS DR13 data set, our group finder identifies 421,797 galaxy groups with halo masses ranging from approximately $10^{11}$ $h^{-1}$Me to $10^{15}$ $h^{-1}$Me. Of these, 62,471 groups contain more than one member galaxy. In Figure 5, we present the distributions of these groups with respect to group richness (i.e.,

number of member galaxies), redshift of the group center, and assigned halo mass. These results are compared against the SDSS group catalog from Lim et al. (2017) (hereafter Lim17 catalog). Despite employing different methodologies for group identification, our results exhibit strong consistency with those of Lim17 catalog across all three distributions. However, our catalog includes a slightly higher number of high-richness groups, particularly at lower redshifts, which leads to a relatively lower abundance of groups at $z < 0.05$. Additionally, our predicted catalog contains fewer massive groups with M_{vir} $10^{15}$ $h^{-1}$Me.

[Figure 5: see original paper]. Number distributions of galaxy groups as functions of group richness (left panel), redshift of the group center (middle panel), and predicted halo mass (right panel). Results from our group finder are shown as blue histograms, while green histograms represent the Lim17 catalog for comparison. The two catalogs exhibit good agreement across all three distributions, with minor differences observed in the abundance of the most massive groups.

Figure 6 further illustrates this through a comparison of halo mass assignments between the two catalogs for matched groups. Group matching is performed using the central and member matching criteria described in Section 4, resulting in 390,177 matched groups. Although the scatter is larger than that observed in the mock catalog predictions, the two catalogs show good agreement in halo mass estimates for low-mass groups. However, our catalog clearly underestimates the halo masses of the most massive groups, resulting in a lower predicted abundance of most massive halos and the divergence in halo mass distribution between the two catalogs. This difference may also be influenced by the underlying simulations used to generate the mock catalogs and estimate the halo mass. Lim et al. (2017) utilized the Evolution and Assembly of GaLaxies and their Environments (EAGLE; Schaye et al. 2015) simulation to calibrate the halo mass function used in their abundance matching. In contrast, our model was trained on data sets from the MS, which adopts a different cosmology. Moreover, due to the data-driven nature of our approach, the model's performance may be influenced by the specific galaxy formation physics implemented in the training simulation.

[Figure 6: see original paper]. Comparison of halo masses for matched groups between our catalog and Lim17 catalog. The x-axis represents the halo masses assigned by Lim17 catalog, while the y-axis shows the corresponding virial masses predicted by our model. For most low-mass groups, the two catalogs yield comparable results. However, our predictions tend to be slightly lower for the most massive groups, consistent with the trend observed in Figure 5.

We further compare our catalog with Lim17 catalog at both the group level and the member (inner group) level. Figure 7 presents the group-level similarity, characterized by the fraction of common matched groups relative to each catalog, denoted as R_c,ML and R_c,Lim, respectively. For groups with halo masses M_{vir} $> 10^{12}$ $h^{-1}$Me, two methods achieve over 90% similarity in both metrics. As previously noted, our model predicts fewer low-mass groups at $z < 0.05$, which contributes to the observed decline in R_c,Lim below $10^{12}$

$h^{-1}$Me.

[Figure 7: see original paper]. Group-level similarity between our catalog and Lim17 catalog. Similarity is defined using the fraction of matched groups relative to each catalog: R_c,ML (green), the ratio of matched groups to the total number of groups in our machine-learning-based catalog, and R_c,Lim (blue), the ratio of matched groups to those in the Lim17 catalog. High similarity (>90%) is maintained for halo masses above $10^{12}$ $h^{-1}$Me, demonstrating strong consistency in group identification between the two catalogs.

We also assess group similarity at the member galaxy level, shown in Figure 8, which displays the cumulative distributions of member-level similarity r_c, defined as the fraction of common galaxies within each matched group. We denote r_c,ML and r_c,Lim as the ratios with respect to the member lists in our catalog and in Lim17 catalog, respectively. For groups with M_{vir} < $10^{14}$ $h^{-1}$Me, over 80% of matched groups exhibit identical member galaxy assignments in both catalogs. For more massive groups, this proportion decreases to approximately 60%, likely due to the increased complexity and richness of higher-mass systems. Overall, these comparisons demonstrate strong agreement between our catalog and Lim17 catalog, both in group identification and in the assignment of member galaxies.

[Figure 8: see original paper]. Cumulative distributions of member-level similarity between our catalog and Lim17 catalog. The similarity ratio r_c is defined as the fraction of shared member galaxies within each matched group. The left panel shows r_c,Lim, the ratio with respect to our machine-learning-based group catalog, while the right panel shows r_c,ML, the ratio with respect to the Lim17 catalog. Results are shown for four halo mass bins, each indicated by a different color. The plots indicate strong agreement in member galaxy assignments, especially for groups with M_{vir} < $10^{14}$ $h^{-1}$Me.

We present the relationship between total stellar mass and virial mass of galaxy groups in Figure 9, subdivided into three redshift intervals: [0, 0.07], [0.07, 0.14], and [0.14, 0.2]. In the figure, solid lines indicate the mean total stellar mass within bins of virial mass, while the shaded regions denote the $1\sigma$ scatter around the mean. Across all redshift bins, groups with M_{vir} < $10^{13}$ $h^{-1}$Me exhibit consistent stellar-to-halo mass relations. As expected for magnitude-limited samples, minimum detectable halo mass increases with redshift due to observational limits. Additionally, deviations among the relations become apparent at higher halo masses, particularly in the z [0.14, 0.2] bin, due to the decreasing number of detectable satellite galaxies at greater distances. These variations highlight that, while stellar mass is a key input to our halo mass estimation model, the network effectively leverages additional group properties, such as richness and satellite spatial distribution, to improve mass predictions.

[Figure 9: see original paper]. Relationship between total stellar mass and predicted halo mass of galaxy groups in three redshift bins: [0, 0.07], [0.07, 0.14], and [0.14, 0.2] in our SDSS group catalogs. Solid lines represent the mean

total stellar mass within bins of halo mass, while shaded regions denote the $1\sigma$ scatter. The relation remains consistent across redshift for groups with M_{vir} $< 10^{13}$ $h^{-1}$Me, with variations at higher masses due to observational limits and reduced satellite detection at higher redshifts.

As demonstrated in Paper I, our model generalizes well to galaxy samples with varying magnitude limits, even when trained on data sets with a flux limit completely different from that in the test sample. To further validate this capability, we manually selected a subset of the SDSS data set by applying a stricter r-band magnitude limit of 17. This selection yields 236,316 galaxies, almost half of the total sample, on which our model identifies 176,822 groups. For comparison, applying the model to the full SDSS data set results in 170,297 identified groups with m_r $<$ 17. In Figure 10, we present distributions of group richness, redshift, and halo mass for the m_r $<$ 17 subset. These are compared with the distributions derived from the full SDSS data set, restricted to galaxies that also meet the m_r $<$ 17 criterion, labeled as m_r $<$ 17.7 in the figure. The distributions show strong consistency between the two data sets, indicating that our model yields stable and reliable outputs across different magnitude limits. However, the virial mass distributions at the high-mass end reveal some discrepancies. This behavior is consistent with our earlier findings in Paper I, which is due to halo mass estimates exhibiting a slight dependence on group richness. While further calibration could mitigate this effect, we emphasize that it does not significantly impact the accuracy of member galaxy assignments.

[Figure 10: see original paper]. Distributions of galaxy groups identified in the magnitude-limited subset (mag_r $<$ 17) compared with those from the full SDSS data set (mag_r $<$ 17.7) restricted to the same subset. Left panel: Number of member galaxies per group. Middle panel: Redshift distribution of group centers. Right panel: Predicted halo mass distribution. Blue histograms represent results from the mag_r $<$ 17 subset, while green histograms correspond to groups in the mag_r $<$ 17.7 data set within the same magnitude range.

Figure 11 displays a one-to-one comparison of predicted halo masses between the two data sets. The results show excellent agreement at lower mass ranges. However, for massive groups, the model applied to the m_r $<$ 17 subset tends to underestimate halo masses, a consequence of the model's sensitivity to the number of member galaxies. This finding underscores the robustness of our method for group identification, while also highlighting the potential need for refinement in halo mass estimation model.

[Figure 11: see original paper]. Comparison of predicted halo masses for common groups identified in the mag_r $<$ 17 and mag_r $<$ 17.7 subsets of the SDSS data set. The x-axis shows halo masses predicted using the full sample (mag_r $<$ 17.7), while the y-axis shows halo masses predicted from the magnitude-limited subset (mag_r $<$ 17). Results are presented as a 2D histogram, where the color in each bin indicates the logarithm of the number of groups. The plot demonstrates strong agreement at lower halo masses, with slight underestimation for massive groups in the mag_r $<$ 17 subset due to reduced group richness.

### 5.1. Catalog Contents

The group catalog constructed by our machine-learning-based group finder and the essential observable properties of galaxies are available at https://github.com/JuntaoMa/SDSS-DR13-group-catalog.git. The contents of catalog data are listed below:

- Column (1) galaxyId: unique ID of galaxies
- Column (2) centerId: galaxyId of its central galaxy
- Column (3) RA: right ascension in degrees
- Column (4) Dec: declination in degrees
- Column (5) z: redshift of galaxy
- Column (6) stellarMass: log(M_*/M_ )
- Column (7) centralMvir: log(M_{vir}/M_ ), where M_{vir} is the virial mass of group of which the galaxy is a member

## 6. Conclusion and Discussion

In this study, we extend our machine-learning-based galaxy group finder from simulation environments to redshift space and real observational data, specifically applying it to the SDSS sample in Lim et al. (2017). We retain the core architecture introduced in our previous work, with minor modifications to enhance compatibility with the observational data set. The model consists of three key components:

1. **Central Galaxy Identifier**. Predicts the central galaxy of a target system using the photometric and spatial properties of the galaxy and its ten nearest neighbors.
2. **Group Mass Estimator**. Predicts the halo mass (M_{vir}) of each group based on the properties of the central galaxy and the five most massive satellite galaxies.
3. **Group Finder**. An iterative procedure that integrates the outputs of the neural networks to produce the final group catalog.

Compared to Paper I, we refined the Group Mass Estimator by excluding the outermost 50% of member galaxies, based on projected distance to the central galaxy, when selecting the top five satellite galaxies. Additionally, we replaced one of the model's inputs, the maximum r-band absolute magnitude among group members (M_r,max), with the redshift of the central galaxy. These adjustments aim to mitigate observational uncertainties and enhance the robustness of halo mass predictions. Furthermore, we train the models on new training data set constructed from MS to better fit the SDSS data.

The model was trained on mock catalog, applying an r-band apparent magnitude limit of 17.7 and restricting redshifts to $z \leq 0.2$. This yielded a training data set of 875,520 galaxies. We evaluated the performance of our group finder using four independent mock catalogs. The model achieved over 90% completeness and purity for groups with halo masses as low as M_{vir}  $10^{11}$ $h^{-1}$Me,

with performance improving further for more massive groups. At the member galaxy level, more than 80% of low-mass groups achieved perfect membership assignment, while over 80% of high-mass groups maintained f_c and f_p values above 0.6 despite the complexities of larger group structures and redshift distortions. Our halo mass predictions were consistent with the true mass distribution, with only minor underestimation at the high-mass end, and could be further improved by potentially incorporating environmental information.

Applying our group finder to the galaxy samples from SDSS DR13, we identified over 420,000 galaxy groups and estimated their halo masses using a trained ANN model. Our results show strong consistency with Lim17 catalog generated by traditional methods, particularly in terms of group abundance, redshift distribution, and halo mass distribution. Comparisons at both the group level and member galaxy level reveal high similarity, with over 90% of groups with M_{vir} $> 10^{12}$ $h^{-1}$Me matched between the two catalogs. Member galaxy assignments also show substantial agreement with Lim17 catalog, achieving exact matches for over 80% of groups in lower mass ranges.

Discrepancies in high-mass halo abundance are attributed to methodological differences in halo mass estimation, where the two works utilized simulations with distinct cosmology and galaxy formation models. Recent efforts have sought to enhance halo mass prediction in group catalogs. For example, Zhao et al. (2025) employed machine learning techniques to mitigate systematic biases between groups with star-forming (blue) and passive (red) central galaxies, achieving halo mass predictions approximately one-third more accurate than those derived from abundance matching. Similarly, Zhang et al. (2024) investigated detailed scaling relations between halo mass and a variety of central galaxy properties, offering improved empirical proxies for group halo mass. In future work, we aim to compare our model's output with these results to assess whether our method already captures such effects implicitly and to evaluate the performance of our prediction with other halo mass proxies.

Additionally, we demonstrated the model's robustness across different observational limits by applying it to a magnitude-limited subset (m_r $< 17$) of the SDSS data set. The model maintained consistent performance, highlighting its generalizability without requiring retraining or additional calibration.

The group catalog, developed using our scheme, along with key observable properties of galaxies, can be accessed at https://github.com/JuntaoMa/SDSS-DR13-group-catalog.git. This repository offers a valuable supplementary resource for numerous research endeavors, particularly those focusing on large-scale structure and the relationship between halos and galaxies.

This research highlights the efficacy and dependability of our deep learning method for identifying galaxy groups and estimating halo mass using actual survey data. The method's adaptability, scalability, and precision position it as an invaluable resource for various extensive surveys, including the 2MASS survey, DESI surveys, along with additional forthcoming surveys. This enables

comprehensive analyses of cosmic structures and the evolution of galaxies.

## Acknowledgments

We acknowledge the stimulating discussions with Jiang Z., Guo Y. and Gao C. This work was supported by the National Key R&D Program of China (2022YFA1602901), the National Natural Science Foundation of China (NSFC, grant Nos. 11988101, 11873051, 12125302, and 11903043), CAS Project for Young Scientists in Basic Research (grant No. YSBR-062), the China Manned Space Program (grant Nos. CMS-CSST-2025-A03 and CMS-CSST-2025-A10), and the K.C. Wong Education Foundation. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

**ORCID iDs** - Juntao Ma https://orcid.org/0009-0000-5513-4100 - Jie Wang https://orcid.org/0000-0002-9937-2351

## References

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543 Albareti, F. D., Allende Prieto, C., Almeida, A., et al. 2017, ApJS, 233, 25 Arbib, M. A. 1995, The Handbook of Brain Theory and Neural Networks, 3361 (Cambridge, MA: The MIT Press) Bell, E. F., McIntosh, D. H., Katz, N., & Weinberg, M. D. 2003, ApJS, 149, 289 Berlind, A. A., Frieman, J., Weinberg, D. H., et al. 2006, ApJS, 167, 1 Bilicki, M., Jarrett, T. H., Peacock, J. A., Cluver, M. E., & Steward, L. 2014, ApJS, 210, 9 Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. 2017, ISPM, 34, 18 Chen, H., Wang, J., Mao, T., et al. 2024, MNRAS, 532, 3947 Crook, A. C., Huchra, J. P., Martimbeau, N., et al. 2007, ApJ, 655, 790 Eke, V. R., Baugh, C. M., Cole, S., et al. 2004, MNRAS, 348, 866 Finkbeiner, D. P., Schlafly, E. F., Schlegel, D. J., et al. 2016, ApJ, 822, 66 Goto, T. 2005, MNRAS, 359, 1415 Guo, Q., White, S., Boylan-Kolchin, M., et al. 2011, MNRAS, 413, 101 Henriques, B. M. B., White, S. D. M., Thomas, P. A., et al. 2015, MNRAS, 451, 2663 Hochreiter, S., & Schmidhuber, J. 1997, Neural Computation, 9, 1735 Huchra, J. P., & Geller, M. J. 1982, ApJ, 257, 423 Knobel, C., Lilly, S. J., Iovino, A., et al. 2009, ApJ, 697, 1842 Lavaux, G., & Hudson, M. J. 2011, MNRAS, 416, 2840 Lim, S. H., Mo, H. J., Lu, Y., Wang, H., & Yang, X. 2017, MNRAS, 470, 2982 Ma, J., Wang, J., Mao, T., et al. 2025, arXiv:2504.01131 Mao, T.-X., Wang, J., Li, B., et al. 2021, MNRAS, 501, 1499 Miller, C. J., Nichol, R. C., Reichart, D., et al. 2005, AJ, 130, 968 Schaye, J., Crain, R. A., Bower, R. G., et al. 2015, MNRAS, 446, 521 Springel, V., White, S. D. M., Jenkins, A., et al. 2005, Natur, 435, 629 Tago, E., Einasto, J., Saar, E., et al. 2006, AN, 327, 365 Tully, R. B. 2015, AJ, 149, 171 Wang, K., Mo, H. J., Li, C., Meng, J., & Chen, Y. 2020, MNRAS, 499, 203 Yang, X., Mo, H. J., van den Bosch, F. C., & Jing, Y. P. 2005, MNRAS, 356, 1293 Yang, X., Mo, H. J., van den Bosch, F. C., et al. 2007, ApJ, 671, 153 Yang, X., Xu, H., He, M., et al. 2021, ApJ, 909, 143 Zhang, Z., Wang, H., Luo, W.,

et al. 2024, ApJ, 960, 71 Zhao, D., Peng, Y., Jing, Y., et al. 2025, ApJ, 979, 42

*Note: Figure translations are in progress. See original paper for figures.*